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　３ 年　 ５ 月 ２６ 日現在

機関番号：１４４０１
研究種目：基盤研究(C)（一般）
研究期間：2017～2020
課題番号：１７Ｋ００２３５

研究課題名（和文）情報セキュリティレベルの高いサーバ・クライアント型メディア認識機構の開発

研究課題名（英文）Development of Client-Server-Based Framework for Privacy-Preserving Media Recognition

研究代表者
中村　和晃（NAKAMURA, KAZUAKI）

大阪大学・工学研究科・助教

研究者番号：１０５８４０４７

交付決定額（研究期間全体）：（直接経費）　　3,600,000円

研究成果の概要（和文）：クライアントユーザから送信されたメディアデータに対しサーバが認識処理を行い，その結果をユーザに返送する，というサーバ・クライアント型メディア認識においては，認識結果や認識モデル，訓練データなどの情報が他者に流出するリスクがある．本研究では，そのリスクがどの程度現実的であるかを分析するとともに，これを防御・回避する技術を検討した．具体的な成果として，サーバにメディア認識結果を把握させることなくユーザにのみ正しい結果を伝送可能な認識機構の開発，サーバの所有する認識器の不正複製物である「認識器クローン」の作成抑止技術ならびに検知技術の開発，認識器からその訓練データを逆推定する技術の開発，を達成した．

研究成果の学術的意義や社会的意義
AI技術の普及によりサーバ・クライアント型メディア認識サービスは既に現実のものとなりつつあり，今後の更なる発展が予想される中で，当該サービスを安心安全に運用・利用できないという事態になれば，大きな社会不安を引き起こす可能性が高い．本研究の成果は，そのリスクを低減するとともに，今後も継続して対処法の研究開発が求められることを示唆するものであり，安心安全なサービスの実現に大きく貢献し得る．また，学術的には，本研究の成果によりメディア認識分野・AI分野と情報セキュリティ分野を融合した新たな研究領域が創出される潜在性を持つ．他形態のメディア認識に対しても同様の研究を行う余地は大きく，極めて意義深い．

研究成果の概要（英文）：Client-server-based media recognition services, where client users send a media data to the recognition server while the server recognizes it and returns the result, have several risks for leaking sensitive information such as the recognition results, the server's recognition model, its training data, and so on. In this research project, we analyzed how much these risks are urgent and proposed some techniques to avoid or defend against them. The outcomes of this research project mainly include (i) media recognition framework where the recognition results are not disclosed to the server but correctly conveyed to the user, (ii) techniques to prevent and detect unauthorized clones of the server's recognition model, which we call "cloned recognizers", and (iii) techniques to estimate and regenerate a training data of a media recognition model only from the model itself.

研究分野：視覚情報処理

キーワード：メディア認識　パターン認識　情報セキュリティ　サーバ・クライアント　認識器クローン　プライバシー保護　Model Inversion Attack

１．研究開始当初の背景

深層学習によるマルチメディア処理技術の進歩とスマートフォンに代表される携帯端末の普及に伴って，近年，サーバ・クライアント型のメディア認識サービスがクラウド上で運用され始めている．その具体的な利用イメージは次の通りである．まず，クライアントユーザは自身の携帯端末からメディアデータを認識サーバに送信する．認識サーバは，独自の訓練データセットに基づいて作成された高度なメディア認識器を有しており，それを用いて，クライアントから送信されたデータを認識する．最後に，認識結果がサーバからクライアントに返送される．この形態のクラウドサービスは，今後，メディア認識サービスにおけるスタンダードとなることが予想される．

一方で，上記のサービスでは，認識対象データの所有者と認識器の所有者が異なることから，次に示す情報セキュリティ上のリスクが存在する可能性がある．

(A) 認識サーバが全てのメディア認識結果を完全に把握できる立場にあるため，認識結果にクライアントユーザのプライバシーに関わる情報が含まれる場合，それがサーバ側に流出する可能性がある．
(B) 悪意あるクライアントユーザが多数のメディアデータを認識サーバに送信し，それらに対応する認識結果（ラベル）を得ることにより，その両者を訓練データセットとして，サーバ側が所有するオリジナルの認識器と同等の機能を持つ複製認識器を作成できる．これにより，オリジナルの認識器が流出した場合と同様の不都合が生じる．

以上 2 点が研究開始当初に想定していたリスクであるが，研究の進捗に伴って，更に次の情報セキュリティリスクの存在が浮上した．

(C) サーバ側が所有する認識器の入出力関係（どのようなメディアデータを入力した時にどのようなラベルが返送されるか）を解析することにより，認識器の作成に使用された訓練データを当該の認識器のみから逆推定できる．これは，非公開の訓練データが流出した場合と同等の問題を生ずる．

２．研究の目的

本研究では，今後より一層の普及が予想されるサーバ・クライアント・サーバ型メディア認識サービスを安心安全に運用・利用できるようにするために，上述の情報セキュリティリスク(A)～(C)がどの程度現実的なものであるのかを分析するとともに，各リスクに対する防御・回避手段を考案することを目的とする．より具体的には，次の個別課題について検討する．

(i) サーバ側にはメディア認識結果が把握できないにも関わらずクライアントユーザにはそれが正しく伝わるような認識機構の実現（リスク(A)の回避手段）
(ii) 悪意あるユーザにより作成された複製認識器（これを本研究では「認識器クローン」と名付けた）の機能が正規サーバ側の所有するオリジナルの認識器と一致しないよう仕向け，その作成を妨げる技術の実現（リスク(B)の回避手段）
(iii) 認識器クローンが作成されてしまった場合に，それを事後的に検知する技術の実現（リスク(B)に対する防御手段）
(iv) 認識器からの訓練データ逆推定がどの程度可能かを明らかにするための逆推定法自体の検討（リスク(C)の可能性の分析）

上記の内，個別課題(i)(ii)は研究開始当初から着手を予定していた課題である．一方，(iii)(iv)は，研究の進捗に伴って新たに浮上した課題である．

３．研究の方法

個別課題(i)～(iv)に対し，それぞれ次の方針で取り組む．

(i)サーバ側による認識結果の把握が困難なメディア認識機構の検討

クライアントユーザは，認識対象のメディアデータ x に対し，その一部を削除・抽象化した改変データ $\tilde{x}$ を生成し，これを元の x と置換した上で認識サーバに送信する，という方略を採用する．この処理により改変データ $\tilde{x}$ には情報に欠損が生じるため，サーバは認識結果を一

意に定めることが困難となる．ただし，このままではユーザもまた正しい認識結果を得られないため，それを可能にする仕組みを検討する．

(ii)認識器クローンの作成を妨げる技術の検討

認識サーバは，認識結果 y を確率 α で別のラベル $\tilde{y}$ に改変し，これを元の y と置換した上でクライアントユーザに返送する，という方略を採用する．上記の方略を悪意あるユーザの視点で見ると，入力のメディアデータ x に対応するラベルが $\tilde{y}$ であるように映る．これは，認識器の入出力関係を誤って理解することに他ならないため，x から本来のラベルである y を認識できるような「正しい」認識器クローンを構築することは困難となる．ただし，上記の方略は正規のユーザに対してはサービス品質の劣化を招く．サービス品質を可能な限り保ちつつ上記の方略を実行する方法を検討する．

(iii)認識器クローンを事後的に検知する技術の検討

映像やプログラムソースコードを対象として盛んに研究されている準同一性検出の考え方を採用する．準同一性検出とは，二つのデータを入力として，一方が他方に軽微な加工を施しただけの複製物（準同一データ）であるか否かを判定する，という処理である．本研究では，正規の認識サーバが，自身の認識器 f のクローンであることが疑われる別の認識器 g に対し，g が真に f のクローンであるか否かを判定する，というシナリオを考える．この判定処理を，(f,g) の組を入力として，準同一性検出の考え方に基づいて実現する．

(iv)認識器からの訓練データ逆推定手法の検討

逆推定攻撃の対象となる認識器が現在主流の畳込みニューラルネットワーク（Convolutional Neural Network; CNN）により作成されていることを前提として，CNN の中核技術である勾配降下法の利用を検討する．具体的には，対象認識器にメディアデータ x を入力したときの出力が所望の値からどの程度乖離しているかを表す損失関数 $L(x)$ を定義し，L を最小化する x を勾配降下法により求めることにより，所望のラベル y に対応するデータを推定する．この時の推定結果は，ラベル y の訓練データに極めて類似すると予想される．また，近年メディアデータの生成技術として注目を集めている敵対的生成ネットワーク（Generative Adversarial Network; GAN）などの深層生成モデルを活用し，逆推定結果の品質向上を試みる．

４．研究成果

(i)サーバ側による認識結果の把握が困難なメディア認識機構の検討

サーバ・クライアント型メディア認識における認識サーバの役割を，認識結果を一意に定めることではなく，認識結果の候補を絞り込むことと捉え直す．その上で，サーバは，「３．研究の方法」の項目で述べたような改変データ $\tilde{x}$ を受け取った際，認識結果の候補を複数列挙し，それらを全てユーザに返送する．一方，ユーザ側では，改変前の元データ x を有していることから，これを Web 空間中に存在する外部情報（外部のメディアデータ）と照合することにより，返送された候補の中から最終的な認識結果を一意に定める．外部情報については，認識サーバ側が認識結果の候補とともにユーザ側に送付する形でも良い．以上の枠組みの概要を図 1 に示す．

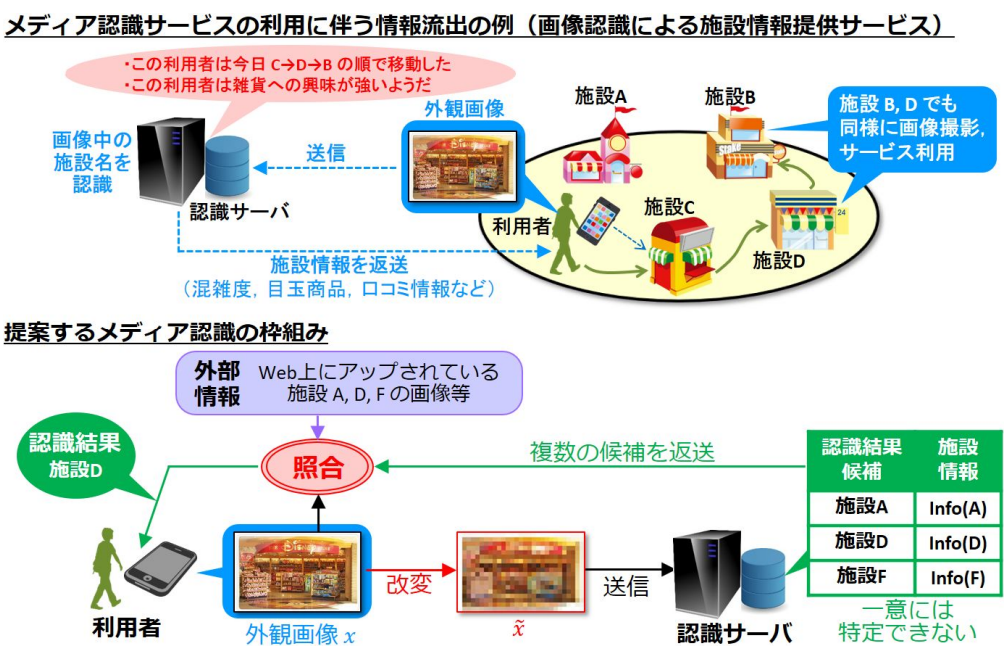

図 1　サーバ側による認識結果の特定が困難なメディア認識の枠組み

提案する枠組みを用いて実際に画像認識実験を試みた結果，サーバ側の認識精度（本来なら認識結果の候補を複数返すべきところで強制的に認識結果を一意に定めた場合の精度）を 41.4%ま

で低減させることに成功した．一方で，ユーザ側での認識精度は 86.9%を維持し，提案する認識機構の有効性が示された．

(ii)認識器クローンの作成を妨げる技術の検討

メディア認識サービスを利用する正規ユーザと，認識器クローンの作成を目論む攻撃者（悪意あるユーザ）とでは，サービス利用時のふるまいに違いが現れることが推測される．後者では，攻撃対象認識器の入出力関係を分析するために多数のデータを認識サーバに送信する可能性が高い．すなわち，サービス利用回数が正規ユーザのそれと比較して非常に大きくなると考えられる．このことを踏まえ，本研究では，「3．研究の方法」の項目で述べたラベル改変のことを「故意誤り」と名付け，その実行確率（故意誤り率）α を各ユーザのサービス利用回数が増えるごとに漸増させることを提案した．正規ユーザの場合，攻撃者と比較してサービス利用回数が相対的に小さいため，α の値は小さいままとなり，サービス品質の劣化をある程度防ぐことが可能となる．一方，攻撃者に対しては，サービス利用回数が非常に大きいことから最終的に α の値も限りなく 1 に近づき，結果として認識器クローンの作成が困難になると期待される．

故意誤りの具体的な方法としては，認識結果を一度求めた後で，そのラベルを直接別のラベルに改変する手法がまず考えられる．このとき，改変後のラベルを一様ランダムに選択する方法の他に，真の認識結果 y と混同しやすいラベルを優先的に選択する方法の 2 種類を検討した．また，別の方法として，クライアントユーザから送信されたメディアデータ x に対し内部的にランダムノイズ ϵ を付加し，$x+\epsilon$ に対し認識処理を実行することにより得られたラベルを改変後ラベルとして使用する方法を検討した．以上 3 種類の方法を用いることにより認識器クローンのクローン精度（認識器クローンと元のオリジナル認識器とで出力ラベルが一致する割合）を実験的に調査した．この結果を図 2 に示す．図 2 より，故意誤りを行わない場合と比較してクローン精度を著しく低下させられていることが分かる．このような認識器クローンの脅威度は低いことから，提案する枠組みにより認識器クローンの作成を一定程度妨げることが可能であることが明らかとなった．

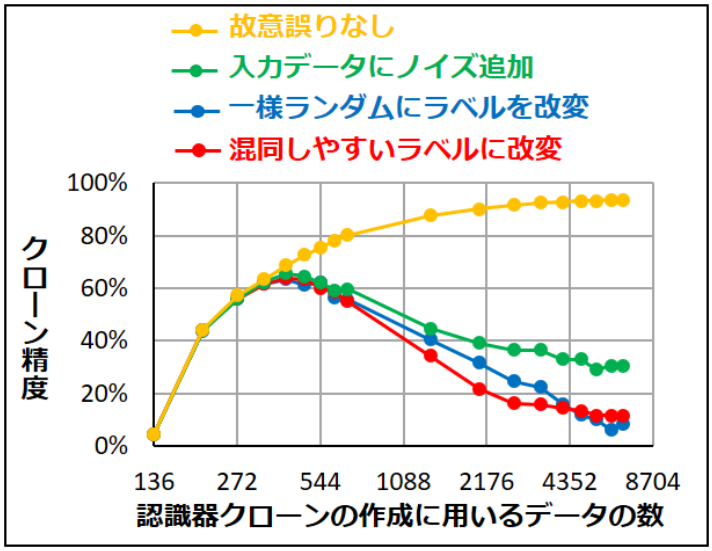

図 2　故意誤りによるクローン精度の低減効果

(iii)認識器クローンを事後的に検知する技術の検討

図 3　認識器クローン作成攻撃とその検知技術の概要

本研究では，「3．研究の方法」の項目で述べた通り，正規のサーバが所有する認識器 f とそのクローンであることが疑われる別の認識器 g の組 (f,g) を入力として，g が真に f のクローンであるか否かを判定する技術を検討する．この処理は f の所有者により実行されるものと想定しているが，このとき，f の所有者にとって g はブラックボックスであることに注意する必要がある．以上の点を考慮し，判定処理用のメディアデータを f と g の双方に入力して各々の出力を取得し，それらを比較することを考える．

一般に，メディア認識器にデータ x を入力した時の出力には，認識結果のラベルに加え，x がどの程度そのラベルらしいかを表す尤度が含まれる．そこで，g が真に f のクローンである

場合とそうでない場合とで，各々が出力するラベルと尤度にどのような違いが現れるかを実験的に調査した．その結果，g が真に f のクローンである場合，両者の出力するラベル同士，および尤度同士が互いに酷似することが分かった．また，尤度に関しては，互いに酷似するだけでなく，g の出力する尤度が f よりも僅かに高くなる傾向があることが分かった．これらは，g が f のクローンでない場合には見られない特性であることから，その有無に基づいて g が f のクローンであることを判定する技術を考案した．その概要を図 3 に示す．

画像認識を対象として提案技術により認識器クローンの検知（判定対象の認識器が真にクローンであるか否かの判定）を試みた実験の結果，87.1%の精度を達成したことから，提案技術の有効性が一定程度示された．

(iv)認識器からの訓練データ逆推定手法の検討

記述を具体化するため，逆推定攻撃の対象は顔画像認識器であるものとし，これを R とする．R は N 人の人物を対象として取り扱うものとすると，顔画像 x が R に入力された時の出力は，x の「i 番目の人物らしさ」を表す尤度（0 以上 1 以下）を i 次元目の要素とするベクトルとなる．これを $y = R(x)$ とおく．ここで，攻撃者が i 番目の人物の訓練データ，すなわち顔画像を逆推定したい場合，y が「i 次元目のみ 1 で，それ以外の全ての次元が 0 のベクトル」となるような x を推定すれば良い．そのようなベクトルを $\hat{y}$ とおく．このとき，「3．研究の方法」の項目で述べた損失関数 $L(x)$ は，例えば

$$L(x) = \|y - \hat{y}\|^2 = \|R(x) - \hat{y}\|^2$$

のように定義できる．上式を最小化する x は，勾配降下法に基づき，適当な $x^{(0)}$ を初期値として

$$x^{(t+1)} = x^{(t)} - \lambda \frac{\partial L}{\partial x}\left(x^{(t)}\right) \quad (t = 0,1,2,\cdots)$$

という処理を反復実行することにより求められる．ただし λ は「学習率」と呼ばれる正の定数である．

本研究では，逆推定結果の品質向上を目的として，上記の手法に対し更に深層生成モデルを追加する．GAN などにより作成される深層生成モデル D は，低次元の乱数ベクトル z からメディアデータ（今回の場合は顔画像）$x = D(z)$ を生成する写像として理解できる．これを用いると，上記の損失関数 L は

$$L(x) = \|R(x) - \hat{y}\|^2 = \left\|R\left(D(z)\right) - \hat{y}\right\|^2$$

のように z の関数として捉えることができる．上式を最小化する z を先程と同様の反復計算により

$$z^{(t+1)} = z^{(t)} - \lambda \frac{\partial L}{\partial z}\left(z^{(t)}\right) \quad (t = 0,1,2,\cdots)$$

として求め，その結果を D に入力することにより，より高品質な逆推定結果を得る．ここで，勾配 $\partial L\left(z^{(t)}\right)/\partial z$ は，R の構造が攻撃者にとって既知であれば誤差逆伝播法により理論値として得る．一方，R の構造が未知の場合は，摂動法により近似値を数値的に求める．

提案手法を実際の顔画像認識器に対し適用することにより得た顔画像の例を，実際の訓練データと対比する形で図 4 に示す．図 4 より，個人の顔の特徴を適切に捉えた画像の推定に成功していることが分かる．また，これらの推定結果を別の顔画像認識器に入力した場合に正しく当人と認識される割合を検証したところ，R の構造が既知である場合と未知である場合の双方において，73%以上の割合で正しく認識された．以上の実験結果から，訓練データの逆推定は現実的に実行可能であり，その対策が急務であることが明らかとなった．

図 4　顔画像認識器からの訓練データ逆推定結果の例

５．主な発表論文等

〔雑誌論文〕　計5件（うち査読付論文　2件／うち国際共著　0件／うちオープンアクセス　2件）

１．著者名 中村和晃，新田直子，馬場口登	４．巻 Vol.30, No.6
２．論文標題 内心プライバシー情報の流出を防ぐ画像認識フレームワークの開発	５．発行年 2019年
３．雑誌名 画像ラボ	６．最初と最後の頁 12-19
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 中村和晃，新田直子，馬場口登	４．巻 Vol.37, No.4
２．論文標題 認識サービスの運用：認識器クローンに対する防御法	５．発行年 2019年
３．雑誌名 MEDICAL IMAGING TECHNOLOGY	６．最初と最後の頁 188-193
掲載論文のDOI（デジタルオブジェクト識別子） 10.11409/mit.37.188	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Nakamura Kazuaki, Nitta Naoko, Babaguchi Noboru	４．巻 14
２．論文標題 Encryption-Free Framework of Privacy-Preserving Image Recognition for Photo-Based Information Services	５．発行年 2019年
３．雑誌名 IEEE Transactions on Information Forensics and Security	６．最初と最後の頁 1264～1279
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/TIFS.2018.2876752	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Mahdi Khosravy, Kazuaki Nakamura, Yuki Hirose, Naoko Nitta, Noboru Babaguchi	４．巻 Vol.15, No.3
２．論文標題 Model Inversion Attack: Analysis under Gray-box Scenario on Deep Learning based Face Recognition System	５．発行年 2021年
３．雑誌名 KSII Transactions on Internet and Information Systems	６．最初と最後の頁 pp.1100-1118
掲載論文のDOI（デジタルオブジェクト識別子） 10.3837/tiis.2021.03.015	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 中村和晃，新田直子，馬場口登	４．巻 Vol.32，No.3
２．論文標題 画像認識サービスの悪用とその対処法に関する基礎検討	５．発行年 2021年
３．雑誌名 画像ラボ	６．最初と最後の頁 27-38
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計9件（うち招待講演　1件／うち国際学会　3件）

１．発表者名 金原祥太，中村和晃，新田直子，馬場口登
２．発表標題 画像認識器に対するクローン訓練攻撃とその防御法に関する考察
３．学会等名 電子情報通信学会パターン認識・メディア理解研究会
４．発表年 2020年

１．発表者名 河津勘介，廣瀬雄基，中村和晃，新田直子，馬場口登
２．発表標題 画像生成ネットワークを用いたModel Inversion Attackの提案
３．学会等名 電子情報通信学会2020年総合大会
４．発表年 2020年

１．発表者名 森勇登，中村和晃，新田直子，馬場口登
２．発表標題 メディア認識サービスにおけるクローン認識器検知手法の検討
３．学会等名 電子情報通信学会2019年総合大会
４．発表年 2019年

１．発表者名
金原祥太，中村和晃，新田直子，馬場口登

２．発表標題
クライアント・サーバ型メディア認識における模倣認識器構築防止手法の検討

３．学会等名
電子情報通信学会2018年総合大会

４．発表年
2018年

１．発表者名
吉村駿佑，中村和晃，新田直子，馬場口登

２．発表標題
構造未知の画像認識器に対するModel Inversion Attackの検討

３．学会等名
電子情報通信学会2021年総合大会

４．発表年
2021年

１．発表者名
中村和晃，森勇登，廣瀬雄基，Mahdi Khosravy，新田直子，馬場口登

２．発表標題
画像認識モデルからの情報流出の可能性とその対処法に関する検討

３．学会等名
第26回画像センシングシンポジウム（SSII2020）（招待講演）

４．発表年
2020年

１．発表者名
Yuto Mori, Kazuaki Nakamura, Naoko Nitta, Noboru Babaguchi

２．発表標題
Detection of Cloned Recognizers: A Defending Method against Recognizer Cloning Attack

３．学会等名
12th Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2020)（国際学会）

４．発表年
2020年

１．発表者名
Mahdi Khosravy, Kazuaki Nakamura, Naoko Nitta, Noboru Babaguchi

２．発表標題
Deep Face Recognizer Privacy Attack: Model Inversion Initialization by a Deep Generative Adversarial Data Space Discriminator

３．学会等名
12th Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2020)（国際学会）

４．発表年
2020年

１．発表者名
Kojiro Fujii, Kazuaki Nakamura, Naoko Nitta, Noboru Babaguchi

２．発表標題
A Framework of Privacy-Preserving Image Recognition for Image-Based Information Services

３．学会等名
23rd International Conference on Multimedia Modeling (MMM2017)（国際学会）

４．発表年
2017年

〔図書〕　計0件

〔産業財産権〕

〔その他〕
画像認識に伴う内心プライバシー情報の漏洩をブロックする仕組みを世界初開発
https://resou.osaka-u.ac.jp/ja/research/2018/20181105_2

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