Vol.2018-MBL-86 No.19
Vol.2018-UBI-57 No.19
2018/2/26

# 発話とその認識エリアに基づく在室管理システムの提案

石澤 鴻弥[1,a)]　岩井 将行[1,b)]

**概要**：現在，在室管理には，高精度カメラ，RFID の NFC や Felica，ジェスチャー，骨格認識，画像認識をはじめとする様々な手法が提案，実用されている．在室管理は，「現在誰がいるのか」「誰がいたのか」を提示する．これにより，コミュニケーションの円滑化や協調作業を支援することにつながる．しかし，一般的に企業で使われることを前提のデータベースなどを利用したシステムの場合，規模が大きく，導入コストがかかるという問題がある．また，そのようなシステムの場合，コワーキングスペースやシェアオフィススペース，研究室のような規模の個々の空間においての導入は，現実的ではない．本研究では，本人の意志に基づいた自然な発話での音声認識をすることによって，特定の空間においてユーザへの負担が少ない在室管理を目標として，在室者の推定や管理を行うシステムを提案する．

# A System of Managing Person Staying in Room Based on Utterance and Recognition Area

Koya Ishizawa[1,a)]　Masayuki Iwai[1,b)]

## 1. はじめに

在室管理には，様々な手法が用いられている．従来の手法としては，紙媒体での在室管理が挙げられるが，IT 化の進む昨今でも東京電機大学の学内にて用いられているのを見かける．手書きで在室を示すものもあるが，紙に予め記載されている教員及び研究室生の名前の表上にマグネットを置くことで在室情報を示す手法が用いられている．他にも，IC カードを用いた在室管理が挙げられる．これら従来の在室管理には，問題がある．それは，自らの手を動かすことによる意識的かつ自発的な操作であるため，手間がかかることや操作忘れが起こり，ユーザへの負担が大きく継続的な利用は困難である．

また，近年では高精度カメラによる自動在室管理をはじめとする，ユーザへの負担がないものがある [1][2]．しかし，そのユーザへの少ない負担と引き換えに，受動的な検知であるため入退室を行った本人の意志とは無関係に判定を行ってしまう．例えば，「トイレ」「飲み物の購入」「ショートミーティング」「気晴らしの離席」をはじめとする退室判定の必要のない短期離席が挙げられる．

現在，シェアオフィススペースやコワーキングスペースを利用したエリアやオフィスが増え始めてきており，そのような環境においての在室管理はユーザにとっては好ましくないといえる．

そこで本研究では，本人の意志に基づいた自然な発話での音声認識をすることによって，特定の空間においてユーザへの負担が少ない在室管理を目標として，在室者の推定や管理を行うシステムを提案する．

### 1.1 Kinect v2 for Windows

Kinect v2 for Windows(以下，Kinect v2) とは，2014 年 10 月に Microsoft 社から発売されたセンサデバイス及び SDK[*1][3] の総称であり，Kinect に搭載されているカメラによる深度情報の取得や人体の認識，マイクアレイによる音声認識といった NUI[*2]技術を Windows の PC で利用することを可能にする．図 1 に本システムで利用した Kinect v2 の外観を示す．

[1] 東京電機大学　未来科学部　情報メディア学科
Kitasenju, Adachi, Tokyo 120–8551, Japan
[a)] koya@cps.im.dendai.ac.jp
[b)] iwai@cps.im.dendai.ac.jp

[*1] Software Development Kit
[*2] Natural User Interface

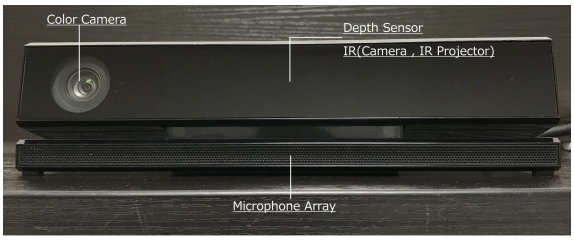


図 1　Kinect for Windows v2

### 1.2　深度センサ

Kinect には深度センサが搭載されており，深度情報[*3]を取得できる．Kinect v2 の深度センサは「Time of Flight（TOF）」という，投光した赤外線が物体で反射して戻る際の時間遅れを測定し，画素単位に深度情報を得る方式が採用されている．また，深度画像の解像度は 512 × 424 であり，取得距離範囲は 500mm〜8000mm である [4](pp64)．外観から深度センサは視認することはできないが，カラーカメラの隣に赤外線カメラとパルス変調された赤外線を投光するプロジェクタが搭載されている．

### 1.3　マイクロフォンアレイ

Kinect v2 にはマイクロフォンアレイ (以下，マイク) が搭載されている．異なる間隔で搭載された 4 つのマイクから構成されているマルチアレイマイクで，水平面音源方向の推定（AudioBeam）や，話者の推定，音声発生源である場所の検出，音声認識，オーディオの録音などを行うことができる．Audio データは，16bit/16kHz のフォーマットでマルチアレイマイクから入力される．

Kinect で音声認識を行うには，Microsoft Speech Platform SDK[5] をインストールする必要がある．また，本システムでは日本語での音声認識を行うため，日本語対応の音響モデルの Kinect for Windows SDK 2.0 Language Packs[6] を使用する．

### 1.4　論文の構成

本論文の構成と各章の概要は以下の通りである．第 2 章で関連研究を論じ，第 3 章では，作成したシステムのアプリケーションの概要や画面，利用方法について説明する．第 4 章では，Kinect v2 の機能の一部である「音声認識」「AudioBeam」「深度情報」を利用した，入退室判定及び個人を特定するための方法を説明し，第 5 章では，本システム及び提案方法の評価実験とその評価結果を考察し，第 6 章では，実験に対してアンケートを取った結果を示し，第 7 章では，本論文のまとめと今後の展望を述べる．

## 2.　関連研究

Kinect を用いた在室管理の研究や人物特定の既存研究として，ジェスチャー認識や骨格情報を利用しているものが挙げられる．

田中らは，利用者があらかじめ登録していたオリジナルのポーズを入退室時にとることで個人を識別し，在室管理を行っている [7]．Kinect for Windows v1(以下，Kinect v1) から取得できる 20 箇所の関節がなす 19 個の角度を，ポーズ登録時とポーズ認識時に求め，それぞれの角度の差分を絶対値で求めている．そして，各関節に設定された係数を差分に掛け合わせた結果，19 個の差分の合計が 180 度以下であれば，同じポーズであると判定している．しかし，利用者は「他人と被らないポーズを考えないとならない」「恥ずかしくないポーズを考えないとならない」とコメントしており，利用者の人数が増加した際の登録するポージングが煩雑化してしまい，十分な精度を得ることが難しくなる可能性がある点や，入退室時の利用者への負担が大きい点から，継続的な利用は難しい．

精度を向上させるために人体特徴量を利用した歩容認識のものとして，横尾らは，Kinect v1 から骨格情報の取得，行動履歴から人物特定を行っている．人物特定の対象となる各人物から学習用データとして多数決機械を作成し，利用時には全てのデータを多数決機械に入力し，判別結果を得ている．行動履歴による人物の判別には，腰の部位骨格情報の X 座標及び Z 座標を使用し，学習用データによる行動履歴ベクトルと利用時の観測される行動履歴ベクトルとのなす角の余弦値から適合度を算出している [8]．評価実験からは，正しく人物の特定が行われているものの，他の被験者として選定された平均評価値が大きいため信頼性に欠ける．また，多数決機械による処理時間が大幅にかかっているため，リアルタイムでの人物特定は困難であり，実用性に欠ける．下久保らは，Kinect から人体の身長特徴量，人体歩行時の関節角度，座標の変異などを特徴量として，サポートベクタマシンとニューラルネットワークを用いて，個人認証するアルゴリズムを定義している [9]．三堀らは，Kienct v2 を複数台を用い，取得したデータを独自のアルゴリズムで統合し個人認証をしている [10]．歩容認識は，事前に学習用データが必要であったりカメラ同様本人の意思関係なしに入退室判定を行ってしまうため，短期離席の判別ができない．

## 3.　アプリケーション

### 3.1　システム概要

システムの概要を図 2 に示す．アプリケーションを立ち上げると，Kinect v2 が起動し，DepthSensor・AudioSensor からそれぞれ Source を取得する．深度情報については，DepthSource から DepthFrameReader を開き，FrameDiscription を取得．それを使用してビットマップやバッファの作成を行う．DepthFrameReader のデータが更新されると DepthFrame を取得でき，そこから実際の深度情報

[*3] センサー面からどれだけ離れているか，という距離情報

Vol.2018-MBL-86 No.19
Vol.2018-UBI-57 No.19
2018/2/26

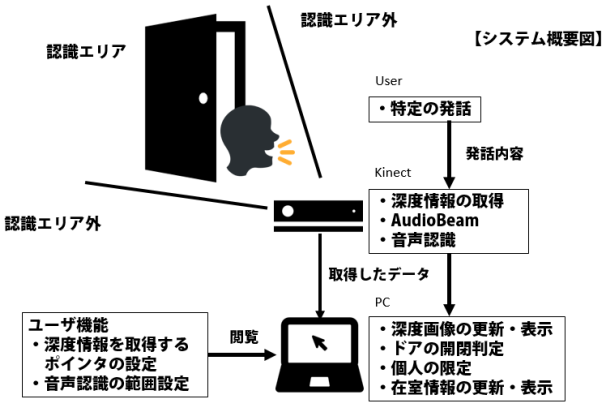


図 2　システム概要図

を取得する．AudioBeam については，AudioSource から AudioBeamFrmaeReader を開き，音声方向を表す BeamAngle を取得する．音声認識については，SpeechRecognitionEngine を作成し，AudioStream を受けるように設定する．また，GrammarBuilder を使用し，SpeedhRecognitionEngine オブジェクトにロードする．

それぞれの初期化処理と認識準備が整うと，研究室メンバーの名前の一覧や深度画像，AudioBeam の設定を行う画面が表示される．

まず，深度情報でドアの開閉を検知するため，ユーザは Depth 画像上のドアをクリックすることで深度情報を取得するポインタを 4 つ設定する．次に，AudioBeam で音声の認識範囲を設定する．そして，ドアが開いており AudioBeam の設定範囲内で音声を認識した場合に，その発話内容から入退室判定及び個人を特定し，画面の表示更新を行う．

### 3.2　アプリケーション画面

アプリケーション画面には，登録された本研究室[*4]の学生名一覧表，深度画像，AudioBeam の設定できるテキストボックス，在室状況のログ，現在の在室人数，アプリケーションを立ち上げてから在室した人数の合計が表示される．図 3 にアプリケーション起動時の状態の画面を示す．

在室状況の確認は，最初は白色である学生名一覧表の各々の背景色が変化することによって一目で把握できる．入室の判定ができた人物の名前の欄は緑色に変化し，退室の判定ができた人物の名前の欄は水色に変化する．表 1 にそれぞれの対応色を示す．

表 1　在室状況の対応色

| 色 | 在室状況 |
|---|---|
| 白 | 不在かつ一度も入室していない |
| 緑 | 在室中 |
| 水色 | 一度は入室していて現在は退室済み |

[*4] 東京電機大学　実空間コンピューティング研究室


図 3　アプリケーショントップ

表 2　入退室の判定を行うために認識する言葉

| | |
|---|---|
| 入室 | おじゃまします　ただいま<br>お疲れ様です　もどりました<br>失礼します　おはようございます<br>こんにちは　おはよう<br>こんばんは |
| 退室 | おじゃましました　あがります<br>お疲れ様でした　またあした<br>失礼しました　さようなら<br>ごはんいってきます |

深度情報を取得するポインタの設定は，画面右に表示されている深度画像内の任意の場所をクリックすることでポインタを設定でき，それぞれのポインタから取得される深度情報からドアの開閉を判定する．

Audiobeam の範囲設定は，画面右にあるテキストボックスを −50°〜+50° の間でそれぞれ数字を入力することで反映される．

## 4.　認識エリアでの発話における入退室判定及び個人の特定

### 4.1　音声認識を用いた発話での入退室判定及び個人の特定

本システムは，入退室時にユーザが特定の言葉に続けて名前を発することで個人判定を行う．システム稼働時は常に音声認識を行っているため，日常会話の中で発せられた名前に反応しないよう表 2 の言葉が発せられた直後のみ判定を行う．

表 2 に本システムで登録した，入退室判定を行う言葉を示す．

表 2 にあるような自然な発話内容での入退室判定を行い，そのあと自分の名前を発することによって個人の特定を行う．在室状況の変化は音声で行われるため，ユーザ自らが手を動かすが必要なく，且つ日常的な発話内容であることからコミュニケーションの活性化につながり，継続的な利用が可能である．

登録されていない言葉の場合は認識されないまた，登録された入退室の言葉のあとに登録された名前を発する必要があり，手順がある．しかし，入退室の言葉を発したあと

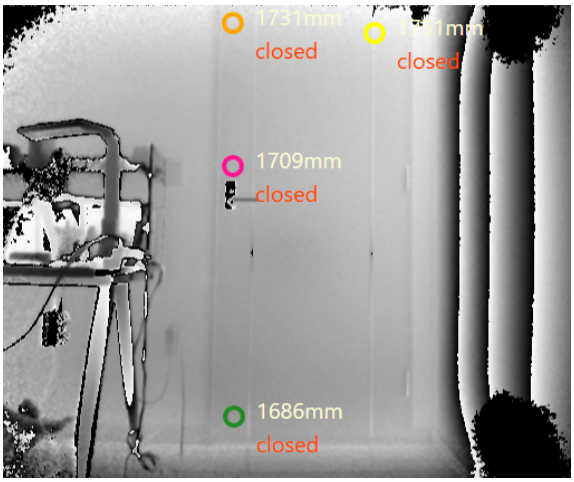


図 4　Depth 画像上での 4 つのポインタ

に自分の名前を発さなければならないという条件であると，複数人のユーザが連続的に入室または退室をした場合において，入退室ごとに言葉を発するのは，テンポが悪く手間であるため，ユーザへの負担がかかってしまう．したがって，入退室の言葉を Kinect v2 が認識したあとの入退室のフラグは，次の入退室の言葉を認識するまで不変である必要がある．

### 4.2　AudioBeam を用いた認識エリアの限定

AudioBeam とは，水平面音源方向推定のことであり，マイクから取得される．取得できる音源方向は Kinect の中心正面を 0° として水平面方向左右に −50°〜+50° の範囲である．

AudioBeam の検出範囲をユーザが画面上で設定することで，例えばドア付近からのみの音声を認識し，ドア付近以外からの音声は認識しないというように，音源方向を限定できるため，認識されるような言葉が研究室内で発せられた場合においてもそれは除外される．

また，Kinect 自体をドアの正面に設置できず，ドアの横になら設置ができるといった設置環境が限られる場合であったとしても，音源方向をユーザが設置環境に合わせて設定することによって，高い汎用性を可能にする．

つまり，設置環境に合わせてより正確に音声認識を行うことができる．

### 4.3　深度情報を用いたドアの開閉検知

深度センサから得られる深度情報を使用し，ドアの開閉を検知する．図 4 に示すように，ユーザが画面上に表示される Depth 画像上の任意の位置を 4 カ所クリックすることで，クリックした位置の深度情報を取得でき，その深度情報の増減を利用している．

取得方法は 2 つ挙げられる． 1 つ目は閉められているドアの面をクリックし，深度情報を取得するポインタを設定

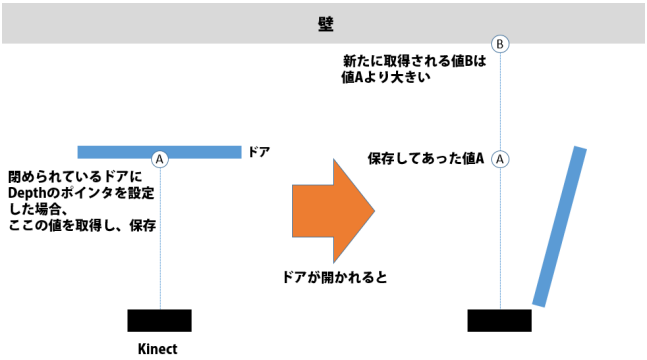


図 5　閉められているドアの面に対して Depth データを取得するポインタを設定した場合

した場合，そのドアが開かれるとそのポインタが取得する深度情報の値はドアの向こう側にある壁の値を取得することになるため，深度情報の値は大きくなる．つまり，ポインタを設定した部分の深度情報の値を一時的に保存しておき，その値が大きくなったら「ドアは開かれている」と判定を行う．図 5 に閉められているドアの面に対して深度情報を取得するポインタを設定した場合の概要を示す．

2 つ目は，開かれているドアの向こう側の壁をクリックし，深度情報を取得するポインタを設定した場合，そのドアが閉められるとそのポインタが取得する深度情報の値はそのドアの面の値を取得することになるため，深度情報の値は小さくなり，「ドアは閉められている」という判定になる

しかし，後者の場合には問題が生じる．ドアが開きっぱなしの状態の場合，ドアの前の廊下または研究室内で Kinect v2 の計測線上を誰かが通り，設定したポインタと被ってしまった場合，深度情報の値は小さくなるため，Kinect v2 は「入室」と認識し False Negative の判定をしてしまう．したがって，前者を用いた前提に設計を行った．

## 5.　評価実験

### 5.1　実験方法

システムを評価するために，本研究室の学生男性 5 人を被験者として，実験を行った．検証項目は，ドアの開閉判定，入退室判定，個人の特定の精度である．実験開始前に，システムの説明と，手順を説明したあとに，順番は特に決めず 1 人ずつ行った． 1 人ずつ入室し，Kinect v2 に認識され，正しく「入室」または「退室」の判定になるまで発声した．同時に，深度情報を取得するポインタの開閉状態を見て，ドアの開閉判定が正しく行われているか確認することで，ドアの開閉判定と，発声の入退室判定のどちらが正しく認識されていないのか判断した．

5 人のうち 4 人が 2 名分の名前を担当し，残りの 1 人が 1 名分の名前を担当することで，合計 9 名分の検証を行った．5 回の実験をし，入室と退室合わせて，90 回の試行を

表 3　正しい名前で認識されるまでに発声した回数

| | 1回目 | | 2回目 | | 3回目 | | 4回目 | | 5回目 | |
|---|---|---|---|---|---|---|---|---|---|---|
| 名前 | 入室 | 退室 | 入室 | 退室 | 入室 | 退室 | 入室 | 退室 | 入室 | 退室 |
| A | 5 | 1 | 3 | 1 | 1 | 1 | 1 | 2 | 2 | 2 |
| B | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 |
| C | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 |
| D | 2 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 |
| E | 4 | 6 | 1 | 1 | 1 | 1 | 2 | 2 | 1 | 6 |
| F | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| G | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| H | 1 | 3 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 3 |
| I | 2 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 |

表 4　実験ごとの平均回数と全体の平均

| 実験 | 状況 | 認識するまでの発話回数の平均 |
|---|---|---|
| 1 回目 | 入室 | 2.11 |
| | 退室 | 1.89 |
| 2 回目 | 入室 | 1.33 |
| | 退室 | 1.11 |
| 3 回目 | 入室 | 1.0 |
| | 退室 | 1.0 |
| 4 回目 | 入室 | 1.33 |
| | 退室 | 1.22 |
| 5 回目 | 入室 | 1.11 |
| | 退室 | 2.0 |
| 全体 | - | 1.41 |

行った．

### 5.2　実験結果

ドアの開閉判定はいずれも正しい結果であった．Kinect センサに人物が近ければ近いほど，その人物が画面に映る割合が多くなるので，深度情報を取得するポインタと被ってしまう．ゆえに，可能な限り Kinect センサから離れたドアの外側から発話することでドアの開閉判定が向上することが分かった．

実験結果を表 3 に示す．

表 4 に実験ごとの平均回数と全体の平均回数を示す．

最大の回数は 6 回であり，その被験者はいずれも E の場合であった．回数が全員 1 回の実験もあった．実験ごとの平均の最大は 2.11 回，全 5 回の実験全体の平均は 1.41 回であった．

しかし，正しく名前を発声しているにも関わらず，別の人物の名前が認識される False Negative の場合があり，実験 1，4，5 において特定の名前に顕著に現れた．E の名前の発声において，名前としては登録済みであるが検証対象ではない 2 名の名前に誤って認識されていることがあった．E の名前の母音は「aai」であり，誤って認識した名前の母音は「aai」「aaai」であるため，母音の流れが同じかもしくは似ている場合に誤認識されると考えられる．

また，言葉の語尾や間に伸ばし棒（ー）を入れた場合にも認識を行っていることも見られた．(例 :「おじゃましまーす」「○○でーす」)

## 6.　アンケート

実験後の 5 人に，アンケートをインタビュー形式で本システムの使用感や利便性について調査を行った．

アンケート項目を以下に示す．

1. 在室状況は把握しやすいか
2. ユーザにとって負担となる点はあるか
3. 操作忘れなど起きず，継続的な利用が考えられるか
4. 問題点はあるか
5. その他感想

アンケート結果を表 5 に示す．

表 5　アンケート結果

| 項目 | 回答 |
|---|---|
| 1 | ・名前の一覧の文字が少し小さいので大きくしてほしい．<br>・在室状況は色だけで分かるのは良い． |
| 2 | ・認識されたかどうかの確認を画面を見なければわからないので，音や光かなにかでなんらかのリアクションがあるとより負担はなくなると思う．<br>・研究室のドアは押さえていないと閉まってくるので，入退室判定されるまで押さえて置けていなければならない点． |
| 3 | ・登録してあった入退室に用いる言葉には普段使うのもあり，その言葉を発した時にふとこのシステムのことを思い出しそう．<br>・言葉だけなので，紙媒体の在室管理と比べると利用でしやすい． |
| 4 | ・「おつかれさまです」は入退室の両方で用いることがあるので間違えそう． |
| 5 | ・声だけで在室の情報が切り替わるのは楽．<br>入退室判定の認識や個人特定の名前認識の登録する言葉を自分自身で自由に登録できると，オリジナルがあっておもしろいかもしれない．<br>・実際に何も言わず入ってくる人や声が小さい人もいるので，このシステムがあると意識的に発話したり Kinect に認識させようと声が大きくなるので良さそう．<br>・声がキーになるため，デバイスやディスプレイに触れる必要がなく，両手が塞がっている場合でも利用できるのが良い． |

## 7.　まとめと今後の展望

本研究では，Kinect v2 の「音声認識」「AudioBeam」「深度情報」を用いた在室管理システムを提案し試作した．能動的ではあるが本人の意志に基づいた低負荷で自然な発話での音声認識や，認識エリアを限定することによってコワーキングスペースやシェアオフィス，研究室といった特定の空間での汎用性やコミュニケーションの活性化を高める．実験の結果，利用時は高い精度でドアの開閉判定と音声認識ができ，在室情報を示すことができた．しかし，名前の母音部分の似た人物の名前で誤って判定されることがあったため，母音が似た名前同士での認識を調査する必要がある．しかし，音声認識の精度はユーザの活舌や Kinect

v2 のデバイスにも依存することから，システム構成の改善も展望として挙げられる．

アンケートからは，言葉だけなのでデバイスやディスプレイに触れる必要がなく，両手が塞がっている場合でも利用できるためユーザへの負担が少ないという意見が得られ，本研究の目的の一部は果たせたと考えられるが，画面を見なくても分かるようなリアクションがあることでより負担を減らせることが分かった．また，操作忘れや継続的な利用に関しての意見は少なかったため，より長期間の運用により，詳細な評価を行っていく必要がある．

## 参考文献

[1] Cnet，http://www.cross-docking.com/service/camera-access/
[2] NEC，http://jpn.nec.com/biometrics/face/index.html
[3] Microsoft, https://developer.microsoft.com/ja-jp/windows/kinect.
[4] 中村薫，杉浦司，高田智広，上田智章，KINECT for Windows SDK プログラミング Kinect for Windows v2 センサー対応版，秀和システム，2015.
[5] Microsoft, Microsoft Speech Platform-Software Development Kit(SDK)(Version 11), https://www.microsoft.com/en-us/download/details.aspx?id=43662.
[6] Microsoft，Kinect for Windows SDK 2.0 Language Packs，https://www.microsoft.com/en-us/download/details.aspx?id=27226.
[7] 田中優斗，福島拓，吉野孝，“入退室時に利用者がとるポーズを用いた在室管理システムの提案”，情報処理学会論文誌，Vol.54，2014.
[8] 横尾亮平，大谷紀子，“Kinect センサによる骨格情報と行動履歴を用いた人物特定”，電子情報通信学会，2013.
[9] 下久保弘樹，北栄輔，“Kinect を用いた歩行動作による個人認証”，情報処理学会研究報告，Vol.2014，No.11，2014
[10] 三堀裕，花泉弘，“Kinect V2 を用いる歩容認識に基づく個人識別手法”，情報処理学会第７９回全国区大会，2017