深度情報を含む映像からの行動認識に関する研究

神園 卓也[1]　高野 茂[2]　馬場 謙介[3]　村上 和彰[4]

概要：大学の遠隔講義等において，受講者の状態を自動的に認識することで，より質の高い講義の実現が期待できる．本研究では受講者の観測手段として，対象までの距離である深度情報の取得が可能であるセンサ Kinect に着目した．深度情報から得られる 3 次元の姿勢情報や部位情報によって，受講者に想定されるいくつかの行動を認識できるかを調査した．

A Study of Human Action Recognition from Video with Depth Information

TAKUYA KAMIZONO[1]　SHIGERU TAKANO[2]　KENSUKE BABA[3]　KAZUAKI MURAKAMI[4]

Abstract: Lectures such as on-line courses of universities are expected to be improved by an activity recognition of students. As a sensor for observation, this paper focuses on Kinext which can capture the depth information of an object. This paper is trying to recognize some kinds of actions supposed for students by 3D-data obtained from the depth information.

1. はじめに

大学における講義や，MOOC(Massive Open Online Course) と呼ばれるオンラインの講義において，講師は講義の内容や自身の講義技術を向上させるために受講者の状態を把握しようとする．受講者の状態とは，受講者が講義に対し，どの程度集中しているか，理解しているかといったことである．講師は受講者の状態に応じて，可能であれば講義中に何らかのフィードバックをする．また，1 人 1 人の受講者に応じたフィードバックを行うことが理想である，しかし，特にオンライン講義において講師が受講者一人ひとりの状態を素早く知ることは困難である．そこで，受講者を何らかの手段でリアルタイムに観測し，観測データから受講者の状態を推定することを考える．これが可能になれば，講師は受講者の観測データを見るだけで，受講者の状態を容易に知ることができる．従って，講師は受講者に対して素早くフィードバックが行えるようになる．また，オンライン講義などにおいて自動で受講者にフィードバックを行うようなシステムも考えられる．本研究の目的は，以上のように，講義中の受講者，特に PC などを用いてオンライン講義を受講している受講者を観測することで当該受講者の状態の推定を容易に行うことができるようにすることである．

本研究では，観測データからどのようにして受講者の状態を推定するか問題になる．そのため，観測データと受講者の状態との関係性を解析することが必要であり，最適な解析方法を調べなければならない．

受講者を観測するセンサとしてカメラ，マイクロフォンや脳波センサ，心拍計といったものが考えられる．文献 [1] における既存研究では，ビデオカメラをセンサとして利用し，受講者の講義中の振る舞いを手作業で分類することで，受講者の講義中の振る舞いと，講義に対する理解度との関係性をある程度明らかにしている．しかし大量の講義中の受講者の振る舞いを手作業で分類するのは困難である．ま

[1] 九州大学工学部電気情報工学科
Department of EECS, School of Engineerring, Kyushu University
[2] 九州大学大学院システム情報科学研究院情報学部門
Department of Infomatics, Faculty of ISEE, Kyushu University
[3] 九州大学附属図書館
Kyushu University Library, Kyushu University
[4] 九州大学大学院システム情報科学研究院情報知能工学部門
Department of Advanced Infomation Technology, Faculty of ISEE, Kyushu University

た文献 [2] における既存研究では講義中の映像から受講者の姿勢を推定し，受講者の状態を推定しようとしている．しかし，受講者の講義中の状態として目を閉じて居眠りをするといった姿勢だけでは推定しづらいものがある．

そこで本研究では受講者を観測するセンサとして Microsoft 社から発売されている Kinect を使用する．その理由として Kinect に搭載されている RGB カメラと赤外線カメラから得られる深度情報を含む映像を用いることで，事前に学習した膨大な人物姿勢パターンから正確な人物の姿勢情報や顔の表情の情報を知ることができ，精度の高い状態推定が期待できるためである．また今後発売が予定されている新型の Kinect ではより精度の高い姿勢や表情認識のほか，心拍数の推定といったことを非観測者が何も身につけることなく実現できる．これも本研究の目的を実現するために Kinect に着目する理由の一つである．

本研究ではまず，Kinect が受講者の状態推定に本当に適しているかどうかを調査する．具体的には簡単な状態として，講義の受講者が取りうる行動を幾つか考える．そして Kinect による行動の認識の可否や方法を調査する．

本稿の構成は以下の通りである．まず第 2 章で，既存研究を紹介する．次に第 3 章で今回使用した Kinect について説明する．そして第 4 章で本研究における行動の推定手法を述べ，第 5 章で今回実施した実験の内容と結果を示す．最後に第 6 章でまとめと今後の予定について述べる．

2. 既存研究

受講者の状態推定に関して，講義の受講者の振る舞いを観察し，受講者の理解度を量ろうとする文献 [1] の研究がある．この研究の目的は，受講者の様子をカメラで撮影して得られた映像データと，授業アンケートや小テストの結果によって，受講者の振る舞いと理解度の関係性を解析することである．この研究では上記の関係性解析においてある程度の成果を挙げている．

また，映像から客観的に観測可能な受講者の振る舞いに注目した研究として文献 [2] や文献 [3]，文献 [2] では受講者の頭，体幹，上腕/前腕の部位状態に着目し，文献 [3] では受講者の反応として「顔上げ」行動に着目している．これらの研究は一定の成果を挙げているが講義中の受講者の振る舞いとして何に着目するかは，一定していない．そこで本研究においては顔や頭の部位姿勢と目の上部と下部の座標に着目する．これによってオンライン講義の受講者の行動として考えられる，集中して画面を見る行動や，居眠りをする行動の認識を目指す．

3. Kinect

本節では本研究で使用する，Microsoft 社の Kinect の概要と Kinect から得られる情報について述べる．

3.1 Kinect の概要

Kinect は Microsoft 社から発売されている RGB カメラや赤外線カメラ，マイクなどのセンサ類で構成されている姿勢計測デバイスである [8]．Kinect は前に立つユーザの頭部，腕，足などの詳細な身体部位の 3 次元位置をリアルタイムで計測できる．具体的には無数の赤外線パターンを照射し，赤外線カメラで被写体に照射されたパターンを観測する．パターンの投影位置から，被写体までの奥行きを計測し人物領域を検出する．その後，事前に学習した膨大な人物姿勢パターンと観測した人物領域を比較し，各身体部位の 3 次元位置を正確に測定する．

3.2 Kinect から得られる情報

Kinect を用いたアプリケーションの開発には Kinect for Windows SDK[9] を用いる．これは Microsoft 社が配布している開発用ライブラリである．Kinect for Windows SDK を用いることで Kinect の各センサから次のような情報を取得することができる．

- Color(RGB 画像)
- Depth(深度)
- Player(人物)
- Skelton(骨格)
- Face(顔の向きやパーツの位置)

3.2.1 Color(RGB 画像)

Color は，1280x960 もしくは 640x480 の解像度で取得することができる．

3.2.2 Depth(深度)

Depth は Kinect が取得することができる情報の中でも特徴的なものであり，Kinect の前に存在する物体との間がどれだけ離れているかを知ることができる．ここで得られる Depth とは Kinect と物体の直線距離ではなく，Kinect に接する平面との距離のことである．図 1 に示す．また Depth を取得する際には Kinect から 0.8[m] の範囲の深度を取得できる「Default Mode」とより近い 0.4[m] の範囲の Depth を取得できる「Near Mode」がある．図 4 に取得できる範囲を示す．得られる Depth は後述する Skelton や Face の認識にも用いられている．本研究では「Near Mode」を用いる．

3.2.3 Player(人物)

Player は Kinect の前に存在する物体が人間であるかどうかを判断することができる．同時に最大 6 人まで認識することができる [5].

3.2.4 Skelton(骨格)

Skelton は膨大な数の姿勢情報を学習した識別器によって，Depth から人物の骨格を取得することができる．骨格の情報として次のような Joint(関節) の 3 次元座標を取得することができる [6].

- 頭

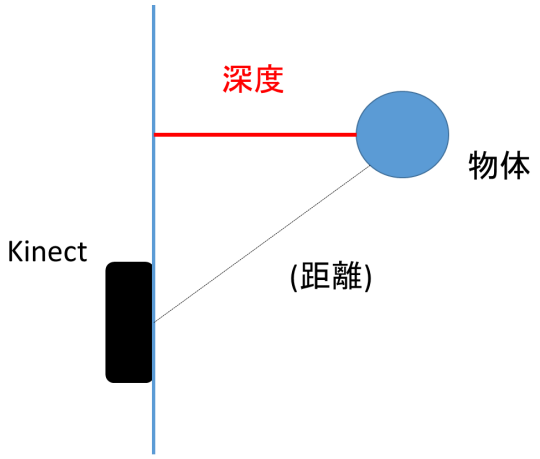

図 1 深度と距離

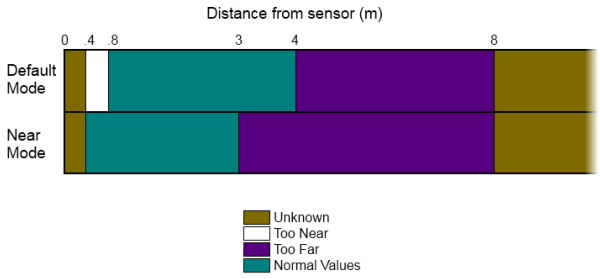

図 2 深度を取得できる範囲 [9]

- 首
- 左右肩
- 左右肘
- 左右手首
- 左右手
- 背
- 尻
- 左右股
- 左右膝
- 左右足首
- 左右足

また Skelton には椅子に座った状態の上半身の骨格を検出する「Seated Mode」がある．このときは次の Joint の 3 次元座標を取得することができる．

- 頭
- 首
- 左右肩
- 左右肘
- 左右手首
- 左右手

Skelton は検出した最大 6 人の Player のうち 2 人まで詳細に Joint を取得することができる．他の 4 人については「尻」の Joint だけ取得することができる．また Kinect は赤外線カメラを中心とする右手座標系が用いられている．図 3 に Kinect の座標系を示す．

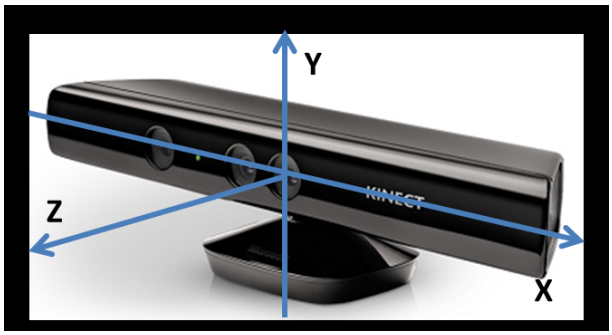

図 3 Kinect の座標系

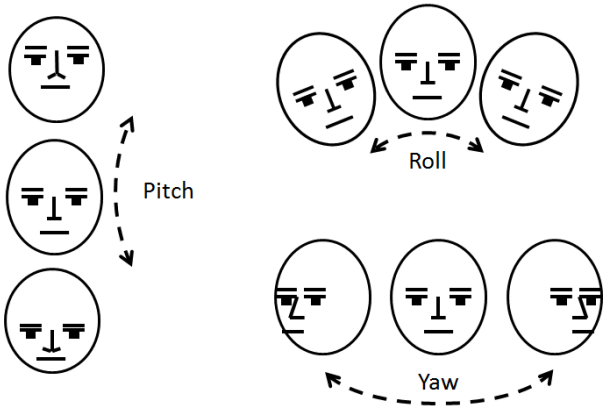

図 4 顔の回転方向 [9]

表 1 Kinect が認識する顔の回転角 [9]

回転軸	値
x 軸 (Pitch)	-90 度=真下を向いている 90 度=真上を向いている ± 20 度以内で認識可能 ± 10 度以内で精度よく認識可能
y 軸 (Yaw)	-90 度=右を向いている 90 度=左を向いている ± 45 度以内で認識可能 ± 30 度以内で精度よく認識可能
z 軸 (Roll)	-90 度=右に傾けている 90 度=左に傾けている ± 90 度以内で認識可能 ± 45 度以内で精度よく認識可能

3.2.5 Face(顔の向きやパーツの位置)

Face では取得した Color や Depth，Skelton を用いて，顔を検出し追跡することができる．Kinect では単純な顔の位置だけではなく，次に示すような顔に関する情報を取得することができる．

- 顔の 3 次元回転角
- 顔の 3 次元平行移動量
- 目や鼻といった顔の 100 箇所のパーツの RGB 画像に位置合わせされた 2 次元座標

図 4，表 1 に顔の回転方向の詳細，図 5 に座標を取得される顔のパーツの詳細を示す．

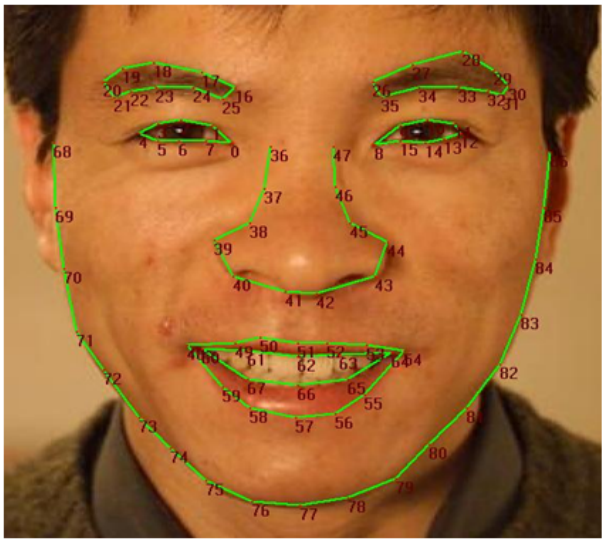

図 5 座標を取得される顔のパーツ [9]

3.3 Kinect データ

3.2 節で述べた Kinect から取得できる情報のうち Color と Depth はセンサから直接得られる，いわゆる「生データ」であるが，それ以外の Player，Skelton，Face は Color と Depth を利用することで得られる高次の情報である．本研究ではこの工事の情報のことを総称して，「Kinect データ」と呼ぶ．Kinect データは人物の姿勢や状況によって取得に失敗することがある．

4. 受講者行動の推定手法

本節では 4.1 節で Kinect から取得したデータから特徴ベクトルを作成する手法について述べ，4.2 で分類する手法について述べる．

4.1 受講者からの特徴抽出手法

4.1.1 利用する特徴量

本研究では Kinect データから以下の 17 個の数値を特徴量として利用する．

- 頭の 3 次元座標
- 左右目の上部と下部の RGB 画像に位置合わせされた 2 次元座標
- 顔の 3 次元回転角
- 顔の 3 次元平行移動量

頭の左右目の部位については図 5 の 2,6,10,14 番に示されている．

4.1.2 Kinect からの特徴量取得

本研究において，Kinect がどのように特徴量を取得するか，図 6 にフローを示す．このフローに従って説明する．まず，Kinect の RGB カメラと赤外線カメラから Color と Depth データを取得する．次に取得した Color と Depth データを用いて特徴量として用いる Kinect データを取得する．取得に成功したら，先程述べた本研究で利用する特徴量を csv ファイルに書き込む．Kinect データは 1 秒間に約 6 回取得することができる．

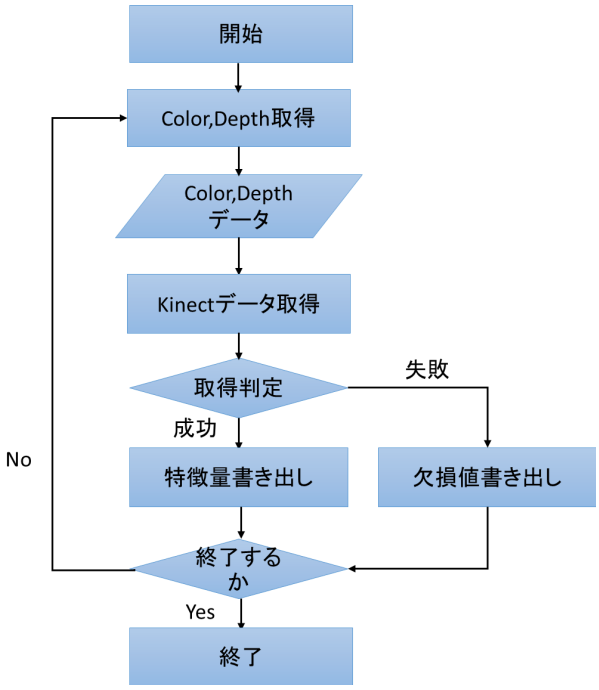

図 6 特徴量を取得するフロー

Kinect データは対象とする人間の姿勢や状況によっては取得に失敗することがある．具体的には顔を下に向けたり，マスクをしたりすると Kinect データの取得が難しくなる．しかし本研究においては受講者に想定される行動として，顔を下にむけて居眠りをするといった Kinect データの取得が難しくなる m のが存在する．よって，Kinect データの取得に失敗した際には各特徴量として「NA(欠損値)」を書き込む．

4.1.3 特徴ベクトルの作成

次に，得られた特徴量からどのように特徴ベクトルを作成するか，図 7 にフローを示す．このフローに従って説明する．まず 1 つの特徴ベクトルを作成するために，連続した 5 つの Kinect データから得られる特徴量を用いる．これは本研究で目的とする「行動」は一瞬一瞬で判断できるものではなくある程度の時間をひとつの単位とすることで判断できるものと考えられるためである．Kinect データは一定の時間で取得できていないが，5 つの Kinect データを用いることで約 0.6 秒間の受講者の動きを 1 つの特徴ベクトルにすることができる．まず 5 つの連続した Kinect データから得られる特徴量がすべて欠損値になっているかどうか判定する．そして 1 つでも特徴量が存在すればそれぞれの特徴量 17 個に対して，平均と分散を算出し，それを特徴ベクトルとして用いる．結果として約 0.6 秒間の行動に対して 1 つの 34 次元の特徴ベクトルができることになる．もし 5 つの kinect データから得られる特徴量がすべて欠損値だった場合には欠損データとする．

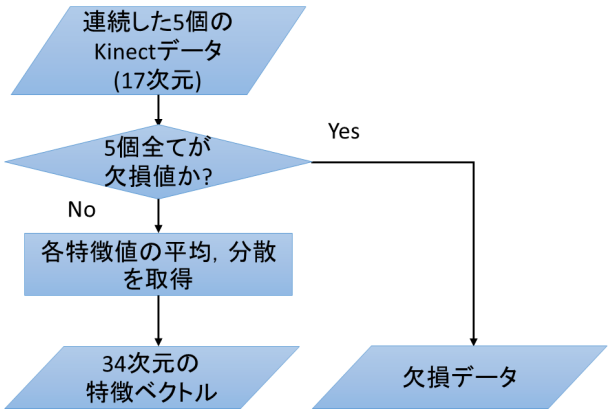

図 7 特徴ベクトルを作成するフロー

4.2 本研究における分類の手順

データ解析を行う際，欠損したデータの取扱いとして補完を行ったり元のサンプルから削除するといったことが考えられる．本研究における欠損データは，その時間に欠損データとなるような行動を取っていた，具体的には頭を下げて居眠りをしていたことなどが考えられる．よって欠損データとなっているものは特定の行動を取っていたと判別する．それ以外の特徴ベクトルに対しては多クラス分類を行うために用いられる k 近傍法を利用した．

4.3 k 近傍法

k 近傍法は教師付き学習アルゴリズムの一つである [4]．非常にシンプルだが，学習データが多くなるほど精度が向上する方法である．

4.3.1 距離

まず，2 データアイテム間の距離を定義する．単なる実数 x_1, x_2, y_1, y_2 を考える．平面上の 2 地点 $(x_1, y_1), (x_2, y_2)$ 間の距離は，

$$\sqrt{(x_1 - x_2)^2 + (y_1 + y_2)^2}$$

これを m 次元のデータ間の距離に一般化すると，

$$\sqrt{\sum_{k=1}^{m}(x_{1k} - x_{2k})^2}$$

つまり，データアイテム x_i と x_j の間との距離 d_{ij} は

$$d_{ij} = \sqrt{\sum_{k=1}^{m}(x_{ik} - x_{jk})^2}$$

と定式化できる．このように定義される距離を，ユークリッド距離と呼ぶ．他にも，マンハッタン距離やマハラノビス距離など，様々な距離がある．以下ではベクトル a, b 間の距離を

$$d(a, b)$$

と表記する．

4.3.2 学習と予測

k 近傍法においては，学習のフェーズでは大した計算は行わず，単に学習データを記憶しておくのみである．この意味で，k 近傍法は，記憶ベース推論や怠惰学習の一種と呼ばれる．以下では本研究で用いる基準変数 Y がカテゴリ変数の場合について説明する．前提として k は設計者が適当に決めておく必要がある．特に $k = 1$ の場合を，最近傍法と呼ぶ．

予測のためのデータ $x \in X$ があると，それに距離が近い学習データ k 個を取り出す．その k 個の学習データの中で最も多い基準変数 Y の値を予測値として採用する．具体的には以下の様な手順となる．

(1) 学習データを予測変数 $X = x_1, x_2, \cdots, x_h, \cdots, x_n$ 基準変数 $Y = y_1, y_2, \cdots, y_h, \cdots, y_n$ とする．x_h は多次元，y_h は 1 次元であるとする．

(2) 予測のためのデータ x が与えられたとする．これから以下のように予測値 y を求める．

(3) $X = x_1, x_2, \cdots, x_h, \cdots, x_n$ のうち $d(x, x_h)$ が小さいものから k 個を取り出し，これを X' と呼ぶ．

(4) X' の取る値のうち，最も出現頻度が多い値を予測値 y とする．

5. 行動推定実験

5.1 実験目的

本研究の最終的な目的は，講義中の受講者を観測したデータを用いて，受講者の状態推定を容易にできるようにすることである．本研究では受講者を観測する手段として Kinect を利用することを述べた．しかし Kinect で得られるデータから人間の状態をどの程度推定できるかわからない．そこで本研究ではまず Kinect から得られる Kinect データから受講者に想定されるいくつかの特定の行動をどの程度認識できるか調査することを目的とする．そのために，特定の行動を定義し，その行動を Kinect で測定する実験を行い，得られたデータを解析することで，Kinect を用いた行動認識の正確さを調査する．

5.2 実験環境

実験はデスクトップパソコンの画面の上に Kinect を設置し，設置した面から約 60cm 離れたところに座って実験を行った．

5.3 実験方法

今回の実験では，著者と同じ研究室の学生 1 人の計 2 人を観測対象とし，PC による講義の受講者に想定されるいくつかの行動を Kinect を用いて測定した．次に観測した

行動の内容を示す．以後は，次に示す行動は括弧内の名称で記述する．

- PC に表示される文章を画面から目をそらさずに読み続ける（集中行動）
- 約 5 秒おきに顔の向きを変えながら画面から目を外す（よそ見行動）
- 頭を完全に下げて居眠りをする（下向き居眠り行動）
- 画面に顔を向け，目を閉じて居眠りをする．（正面居眠り行動）
- 画面から大きく距離を取り（約 80cm から 1m），他人と会話を行う（散漫行動）

この 5 つの行動を 1 回 30 秒間ずつ，Kinect で記録した．著者が 2 回，もう 1 人の学生に 1 回の記録を行った．

得られた Kinect データから 4.1 節で示した方法で特徴ベクトルを作成した．その特徴ベクトルに対してそれぞれの行動の名前をカテゴリ名として付与した．また，欠損データにも同じように行動の名前をカテゴリ名として付与した．しかし判別を行う際には，欠損データはすべて下向き居眠り行動と分類した．これは今回上げた 5 つの行動のうち下向き居眠り行動が Kinect データの取得がほとんど行えず，欠損データならば下向き居眠り行動をとっていると考えられるためである．

欠損データでない，特徴ベクトルに対しては 4.3 節で示した k 近傍法を用いて分類を行った．また $k = 3$ とした．学習用データとテストデータの分類は行わず，1 つのデータをテストデータ，残りを学習用データとしてこれをすべてのデータが 1 回ずつテストデータとなるように繰り返した．

データ数は欠損データが 427 個，特徴ベクトルが 422 個となった．

5.4 実験結果

5.4.1 評価尺度 [7]

複数のカテゴリへの分類評価は，様々な観点からの評価尺度を用いることができる．本研究では従来から用いられる適合率，再現率，およびそれらを組み合わせた F 値を用いた．適合率，再現率，F 値はそれぞれ以下のように求めた．

- 適合率 = 分類した行動の中で正解と一致した数 / ある行動に分類した数
- 再現率 = 分類した行動の中で正解と一致した数 / 実際のある行動の数
- F 値 = (2× 適合率 × 再現率) / （適合率 + 再現率）

5.4.2 結果

図 8 に分類した結果を示す．また図 9 に各行動についての適合率，再現率，F 値を示す．また全体の正解率は 62.54%となった．

正解データ

予測結果	散漫行動	よそ見行動	下向き居眠り行動	集中行動	正面居眠り行動
散漫行動	195	9	0	3	2
よそ見行動	0	92	0	1	7
下向き居眠り行動	9	78	136	102	102
集中行動	0	1	0	51	0
正面居眠り行動	0	4	0	0	57

図 8 実験結果

	適合率(%)	再現率(%)	F値
散漫行動	93.30	95.59	0.9443
よそ見行動	92.00	50.00	0.6479
下向き居眠り行動	31.85	100	0.4831
集中行動	98.08	32.48	0.4880
正面居眠り行動	93.44	33.93	0.4978
平均	81.73	62.40	0.6122

図 9 各行動の適合率，再現率，F 値

5.4.3 考察

一般的に適合率と再現率はトレード・オフの関係にある．しかし，今回の実験のうち，散漫行動に関しては適合率，再現率ともに 90%を超える良い結果が出ている．考えられる要因として，今回の行動のうち散漫行動のみが Kinect センサから大きく離れたものであり，また Kinect の深度計測が正確なものであったことが考えられる．この結果から，講義の受講者が取る行動のうち画面に近づいたり離れたりといった行動はかなり正確に認識できるといえる．下向き居眠り行動については適合率が低く，再現率が高い．これは受講者が下向き居眠り行動を実際に行った時に正確に判断できるが，下向き居眠り行動を取っていない時でもその行動をしていると判断してしまうと言える．その他 3 つの行動に対しては，適合率が高く，再現率が低い．これは先程とは逆に，その行動をとっていると判断した時には正確であるといえるが，実際にその行動をとっている時に見逃していることが多いといえる．

このような結果となった原因として，欠損データの取り扱い方があると考えられる．今回の実験では，欠損データをすべて下向き居眠り行動と分類した．下向き居眠り行動を取れば欠損データが生まれると言えるが，欠損データには他の行動をとっている時のものもかなりの数がある．これを改善するためには，特徴ベクトルを作成するときに用いる Kinect データの数を増やすことですべてが欠損値となりづらいようにするといった方法が考えられる．

今回の実験では Kinect を用いることで画面に近づいた

り離れたりする行動についてはほぼ正確に認識することが分かった．しかしその他の行動については，適合率や再現率に改善の余地があり，またそれは欠損データの取り扱い方にあると考えられる．今後は欠損データをいかに減らすか，調査する必要がある．

6. おわりに

本稿では，講義中の受講者の状態推定を行うことを目的とし，深度情報を含む映像からどの程度の状態推定が可能であるか実験を行った．結果として，頭の位置や向きの変化があるよそ見や居眠り等の行動は，高い適合率で認識できることが明らかになった．

今後は欠損データを減らす特徴ベクトル作製法の検討や，今回試した以外の行動についての実験や別の特徴量を考えた実験を行うことを検討する．

参考文献

[1] 椋木 雅之，上松 信，美濃 導彦: 項目反応理論に基づく理解度と振る舞いの関係性解析, 教育システム情報学会誌, Vol. 30, No. 1, pp. 65-76, (2013).
[2] 椋木 雅之，美濃 導彦: 講義室での受講生の振る舞い観測と理解度推定の研究, 第 26 回 人工知能学会全国大会, pp. 1-4, (2012).
[3] 山根 卓也，中村 和晃，上田 真由美，椋木 雅之，美濃 導彦: 講義中の行動分析に基づく講師受講者間インタラクションの検出, 人工知能学会 先進的学習科学と工学研究会, Vol. 60, pp. 7-14, (2010).
[4] 井上創造: データ科学特論, http://kiso.mns.kyutech.ac.jp/attachments/12524/k-nn.pdf, (参照 2014 年 2 月).
[5] 中村 薫，齋藤 俊太，宮城 英人: KINECT for Windows SDK プログラミング C++編, 秀和システム, (2012).
[6] 杉浦 司，中村 薫: 「Kinect for Windows SDK」実践プログラミング - 「深度センサ」「カラーセンサ」「マイクアレイ」を活用! (I/O BOOKS), 工学社, (2013).
[7] 石田栄美, An-Shou Cheng, Douglas W. Oard, Kenneth R. Fleischmann: 人の価値観を表すカテゴリを対象にした複数カテゴリへの自動分類の試み, 文化情報学 : 駿河台大学文化情報学部紀要, Vol. 16, No.2, pp. 53-68, (2009).
[8] KINECT for Windows, http://www.microsoft.com/en-us/kinectforwindows/, (2014 年 1 月確認).
[9] KINECT for Windows SDK, http://msdn.microsoft.com/en-us/library/hh855347.aspx, (2014 年 1 月確認).