3D アバタによるインタラクティブな Web 講義の実現に向けた Web ブラウザにおける 3 次元骨格情報に基づく講義の再現

竹内 章裕[†1,a)]　長谷川 大[†2]　佐久田 博司[†2]

概要：近年，e-learning にインタラクティブ性を導入するために，html5 や JavaScript が利用されるようになった．また，グラフィックプログラミングを支援する JavaScript API である WebGL の技術により，Web ブラウザ上で人間らしく振る舞う CG キャラクタのアニメーションを表示することが可能となっている．Web 上で利用される講義動画形式の e-learning は，これらの技術の発達に伴い，インタラクティブ性の向上が望まれる．本稿では，モーションキャプチャ装置によって取得された，講義を行う講師のモーションデータを利用した，Web 上で 3D アバタがインタラクティブな講義を行う e-learning システムを提案する．講義動画の代わりに 3DCG の講師アバタを利用することで，静的な講義アーカイブをインタラクティブなコンテンツに変更することが可能となる．本システムの開発を目的に，モーションキャプチャ装置を利用して講義の 3D モーションを撮影し，短い講義データを作成した．また，作成した講義の 3D モーションデータを処理し，Web ブラウザ上で 3D 講師アバタによって講義が行われる視聴システムの開発も行った．

Web Browser Lecture Viewer Based on the 3D Skeleton Information toward Interactive Lecture Performed by 3D Avatar

AKIHIRO TAKEUCHI[†1]　DAI HASEGAWA[†2]
HIROSHI SAKUTA[†2]

Abstract: To introduce interactivity in e-learning, html5 and JavaScript have been widely used for years. And recent advancements of WebGL, JavaScript API for computer graphic programming that works on any modern browsers, enable us to draw animations of a fully human-like embodied computer graphics character on web-browser. With these technologies, web-based e-learning could be more interactive. We propose Interactive Lecture, where lecturers' motions are captured by a motion capture system, then a 3D avatar interactively performs the lectures. The advantage of the use of computer graphics character to perform lectures instead of videos is that enable e-learning designers to replace static lecture to interactive contents. To develop our system, we recorded 3D motions of a lecture performed by a volunteer with a motion capture system, and created a small lecture data. We also developed a web-based viewer that can process the 3D motion data and draw a lecture performed by a 3D avatar on a web-browser.

1. はじめに

近年，高等教育において，学習者自身が必要な学習課題を考え，科目を自由に選択し，学習を行うことができる環境が求められている．特に，工学分野においては，急速な技術の発達に合わせて，学習すべき科目も増加していくことが考えられる．このような需要の下で，Massive Open Online Courses (MOOCs) の重要性が注目されている．

MOOCs では，数多くの実際の授業風景をビデオカメラで撮影し，撮影した講義動画を巨大なデータベースとして蓄積する. 学習者は, 蓄積された大量の講義動画の中から，自分が学びたい科目を自由に選択することができる. また，インターネットを通していつでもどこからでも学習を行うことができるという特徴を持っている．しかし，講義を撮影するだけの講義動画は, アーカイブ作成が容易であるが，講義動画を視聴する学習者にとっては，動画を再生するだけとなり，講師と対面して行われる高いインタラクティブ性のある講義のように，講師が学習者の理解度を確認しながら授業を進めたり，学習者が講師に質問をしたりせず，一方向的な学習となる．このような，動画再生を行うことしかできない e-learning では，学習者の学習意欲の維持や学習効率の向上は困難であると考えられる．

一方で，Web アプリケーション技術の向上により，Web ブラウザにおいて非同期にデータの送受信を行うことや，JavaScript API である WebGL の技術によって，Web ブラウザ上で人間らしい CG キャラクタを表示することが可能となっている．そこで，これらの技術を取り入れた，より高度な e-learning システムの開発が期待されている．

このような理由から，本稿では Web ブラウザ上で 3D アバタが講師役となり，インタラクティブに対面式の講義を行う e-learning システムを提案する．本システムは図 1 で示すように，講師の身体動作データを講義の分節ごとに分割しデータベースに蓄積する．そして，学習者の状況に応じて, 必要なモーションデータをデータベースから探索し，Web ブラウザ上の 3D 講師アバタが，そのモーションを利用してインタラクティブに講義を行う. この仕組みにより，

†1 青山学院大学大学院
Graduate School of Aoyama Gakuin University
†2 青山学院大学
Aoyama Gakuin University

a) c5614134@aoyama.jp

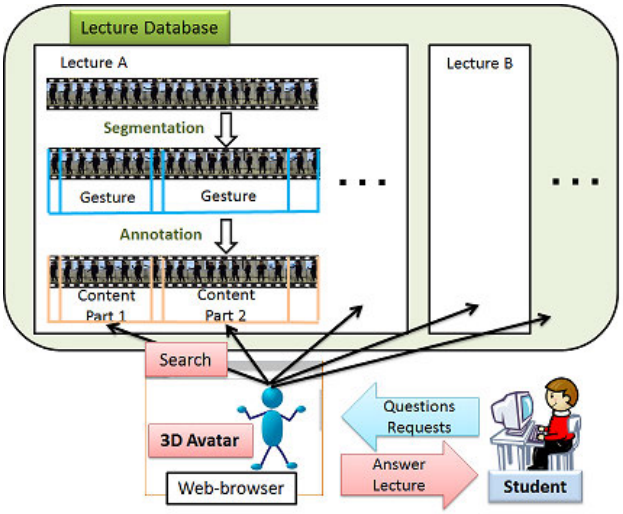

図 1　3D アバタによるインタラクティブな講義システム
Figure 1　Avatar represented interactive lecture system.

e-learning のコンテンツ作成において，学習者の状態や講義内容に応じて，非言語情報を自由に操作することが可能となり，対面型講義における講師のアイコンタクトや表情，ジェスチャなどの非言語情報の伝達が，講師アバタを通して利用可能となる．また，学習者の理解度や講義の内容に応じて，複数のコンテンツを組み合わせることによって，1 体のアバタが多様なコンテンツをカバーする，総合的な学習ガイドとしても利用可能となる．

このシステムのデータベースには，一連の講義のモーションを，容易に再利用ができるように，講師の発言を文節ごとといった意味のある単位ごとに区切り，それに応じたモーションデータを細かく分割し，蓄積しておく（アーカイビング）．このデータベースを人手で作成するのでは，講義データの増加に伴い，膨大な時間やコストが掛かってくるため，自動的なアーカイビングを行うことが求められる．

そこで，講師のジェスチャ動作を認識し，ジェスチャの単位でモーションデータを自動的に分割することを試みた．

以下に，本稿の構成を述べる．2 章では，関連研究について紹介する．3 章では，インタラクティブな講義において必要となる，講義データのアーカイブを行うシステムについて述べる．4 章では，モーションデータを分割する方法について述べる．5 章では，分割システムの精度について，調査した結果を述べる．6 章では，本稿の結論と今後の展望を述べる．

2.　関連研究

インタラクティブな知的教育システムの研究分野において，学習効果を向上させるための手段として，人型のキャラクタ（エージェント）を利用する研究が多数行われている[1][2]．エージェントを利用した教育の研究として，長谷川らは，身体動作を行うエージェントを利用した e-learning を設計し，身体動作を行うことによる教育効果の向上を確認している[3]．また，Baylor らは教育エージェントが行う講義の形式や指示ジェスチャ，感情表現などの非言語情報の効果的な設計手法を提案している[4]．これらのエージェントの設計手法に関する研究以外にも，学習効果を高めるために，エージェントを聞き手として利用する方法[5]やエージェントと一対一の学習環境ではなく，複数人での共同学習環境においてエージェントを利用する方法[6]のようなエージェントの利用方法に関する研究も行われている．

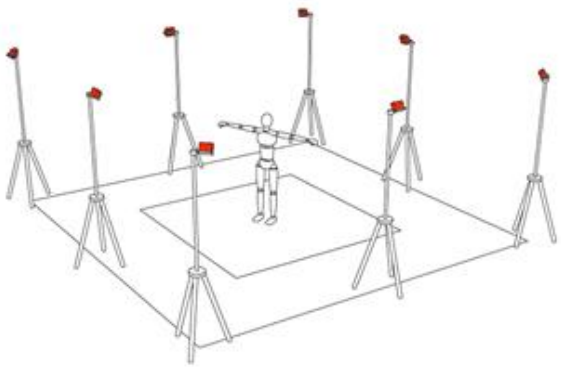

図 3　モーションキャプチャシステムのセッティング
Figure 3　Motion capture system settings.

しかし，これらの研究に用いられるエージェントを利用する教育教材は，新たな学習コンテンツを追加するために，システムも含めて新しく作成する必要があり，学習コンテンツを充実させることが困難である．そこでビカスらは，ハイパーテキスト型映像シナリオ記述言語を利用して学習コンテンツを作成する手法を提案した[7]．しかし，このアプローチは，１つのコンテンツを製作することは比較的容易であるが，多数のコンテンツを作成するには，１つ１つ記述していく必要が存在する．また，ハイパーテキストのタグの種類に応じて表現力が制限されてしまうという問題もある．

このような現状を踏まえ，本稿では一般的に行われている実際の講義をアーカイブし，それをもとに講師のアバタがインタラクティブな講義を行う学習システムを提案し，開発を行った．また，モーションキャプチャ技術が発達し，人間の動作を容易に利用可能となった．そこで，本システムでは，講師の動作も含めてアーカイブを行い，アーカイブしたモーションデータを利用することで，講師アバタの非言語情報を表現する．

3.　システム概要

本章では，作成したシステムの概要について述べる．本システムでは，図 2 で示すように，取得したモーションをデータベースに蓄積し，そのモーションデータをブラウザ

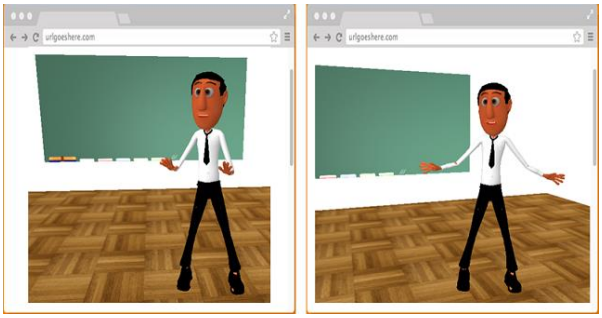

図 4　Web ブラウザ上で講義を行う 3D アバタ
Figure 4　3D Avatar performing a lecture in a web-browser.

上の 3D アバタの動作として利用する．

まず，モーションの取得とデータベースへの蓄積に関しては，モーションキャプチャ装置を利用して，人間の講師が講義している様子を撮影する．今回使用したモーションキャプチャ装置はフレームレートを 120fps に設定したカメラ（OptiTrack Prime 41）8 台を図 3 のように配置し，音声レコーダを講師の前に設置したものとなっている．また，講義を行う講師の体に，合計 49 個の専用マーカを取り付けておく．カメラで撮影したマーカの 3 次元位置座標をソフトウェア（Motive: Body）上で計算を行う．その後，同ソフトウェア上での計算により，51 個の関節を持つ人体骨格情報とフレームごとのすべての関節の回転角度情報を取得することができる．取得したデータは回転角度が Euler 角で定義された Bounding Volume Hierarchy（BVH）形式で出力される．最後に，出力された BVH データを音声データとともに，講師が行うジェスチャ単位で分割し，データベースに蓄積する．

続いて，蓄積されたモーションデータの利用に関して，本システムはクライアント PC の Web ブラウザ上で以下のような動作を行う．まず，システムは講義モーションのデータベースにアクセスを行い，学習者が必要とする講義モーションと音声のデータセットを取得する．同時に，表示する 3D モデルのデータも取得しておく．この講師アバタの 3D モデルは，あらかじめ 3D モデリングソフト（Blender）を利用して，骨格構造をモーションデータ内で定義されている人体骨格情報に合わせてある．そのため，システムが関節の回転を行う際に，モデルの骨格構造を計算することで，3D モデルのアニメーションが行われる．また，通常の PC モニタの最大リフレッシュレートは 60fps となっており，モーションキャプチャシステムで取得された 120fps のモーションデータと異なる．そのため，クライアントの Web ブブラウザでアニメーションの描画を行う際に，モーションデータのフレームレートを減少し，モニタのリフレッシュレートと同様のフレームレートで描画を行っている．Web ブラウザ上での CG 利用においては，JavaScript API である WebGL によって，クライアント PC の GPU を利用して，3D モデルのレンダリング計算が高速に処理される．

我々は，この 3D 講師アバタが，汎用的な PC 上で正常に動作を行うか検証を行い，Web ブラウザ上で頂点数が 9310 個のアバタが，20fps から 50fps で描画されることを確認した (Intel core i7-4500, Intel HD Graphics Family, 8GB RAM)．表示された 3D 講師アバタを図 4 に示す．

4.　ジェスチャ単位による講義データの分割

本システムでは，学習者の状況や要求に応じて，データベースから適切なデータを探索し，3D アバタの動作を，動的に生成する．そのために，講義のモーションデータは発言の内容に基づいて細かく分割し，データベースに蓄積しておく必要がある．発言の内容に基づいて分割する方法として，自然言語処理が考えられるが，音声入力に対する，コンピュータの正確な自然言語処理技術は，非常に難しく，必要とする処理も膨大なものとなってしまう．

そこで，我々は講師の行うジェスチャ動作によって文節の境界を見つけ出す方法を利用することとした．一般的に，ジェスチャは発言に関連して発生するため，ジェスチャと発言の内容は，強く結びついていると考えられている[8][9][10]．これらの理由から，我々は，ジェスチャ動作が，一連の講義データを文節ごとに分割するための手がかりになると考え，講師のジェスチャ動作に基づいた，文節ごとの講義分割を試みた．さらに，モーションデータの利用において，腕の 3 次元位置と回転角度の認識は，音声認識と比較して正確である上，いかなる言語に対しても適応することが可能である．

McNeil によると，ジェスチャは片腕もしくは両腕の肩から指にかけての部位が行う動作として定義される．また，ジェスチャ単位として，ジェスチャを行っていない状態（レストポジション）から，四肢が動き始めた瞬間から，再びレストポジションに到達した瞬間までの区間と定義し，状態に応じて Preparation, stroke, retract, hold の 4 つのパートに分類するとしている[10]．

この定義に基づいて，我々は，ジェスチャ動作を行っているかを判断するために，ある時点での腕の位置とレストポジションの時点での腕の位置との差を利用した．今回は，両肩と両肘の合計 4 つの関節の角度を利用し，特定のフレームとレストポジション時点でのフレームとの間接角度の差を計算し，1 秒間分の平均を判定用の評価値とした．式（1）を利用して，各関節における角度の差分を算出する．

$$d_{j,t} = \frac{\sum_{i=Ft}^{Ft+FPS}(\Delta X_i + \Delta Y_i + \Delta Z_i)}{FPS} \qquad (1)$$

ここで，i は間接の名前，t は判定を行う時間[秒]，i はフレーム番号，Ft は時間 t におけるフレーム番号，ΔX_i, ΔY_i, ΔZ_i は，レストポジション時のフレームとターゲットフレーム i における各軸の関節角度の差を示している．FPS は

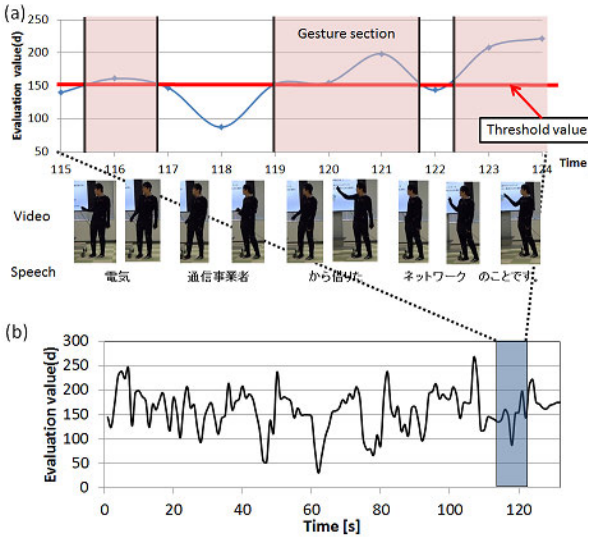

図 5　ジェスチャ動作に基づく講義分割

Figure 5　An example of lecture segmentation based on gesture units.

モーションデータのフレームレートとなっており，本稿では，120fps のモーションデータを利用したため，120 に設定されている．

また，式（2）を利用して，4 つの関節（両肩と両肘）の差分の合計を算出する．

$$d_t = \Sigma_{joints} d_{j,t} \qquad (2)$$

図 5(a)で，これまでに説明したレストポジションとの距離に基づいたジェスチャ単位認識の例を示す. 図において，システムは，先ほどの計算によって求められた評価値d_tがあらかじめ定めておいた閾値よりも高い区間をジェスチャを行っている状態として判定し，評価値d_tが閾値よりも低い区間をジェスチャを行っていない状態として判定する．

5.　分割システムの評価

本システムを評価するために，2 分間の講義を用意した．講義はモーションキャプチャ装置とビデオカメラで同時に撮影が行い，システムによる分割と人手による分割をそれぞれ行い, それぞれの結果を比較した. 人手による分割は，録画した講義のビデオを見て，第一著者が，ジェスチャを行っている状態とジェスチャを行っていない状態のラベル付けを行った．また，今回は直立している状態をレストポジションとして設定した．

図 5(b)に，講義の各時間[秒]において，システムが算出した評価値の時間変化を示す．システムが行ったジェスチャ単位分割の結果を評価するために，ジェスチャの開始タイミングと終了タイミングの精度を計算した．表 1,2 と図 6 で，閾値を変化させた場合のジェスチャユニットの開始タイミングと終了タイミングの適合率，再現率，F 値を示す．人手による分割には，ビデオ視聴ソフトなどによるわずかな誤差が存在するため，前後 1 秒を許容して比較を行った．

表 1　ジェスチャ分割の結果（開始タイミング）

Table 1　Results of gesture segmentation (start point).

Threshold value	120	130	140	150	160	170	180
Precision	0.79	0.76	0.75	0.78	0.67	0.78	0.67
Recall	0.59	0.71	0.71	0.82	0.71	0.71	0.59
F-measure	0.67	0.73	0.73	0.80	0.69	0.74	0.63

表 2　ジェスチャ分割の結果（終了タイミング）

Table 2　Results of gesture segmentation (end point).

Threshold value	120	130	140	150	160	170	180
Precision	0.86	0.82	0.81	0.83	0.78	0.72	0.73
Recall	0.65	0.76	0.76	0.88	0.82	0.76	0.59
F-measure	0.74	0.79	0.79	0.86	0.80	0.74	0.65

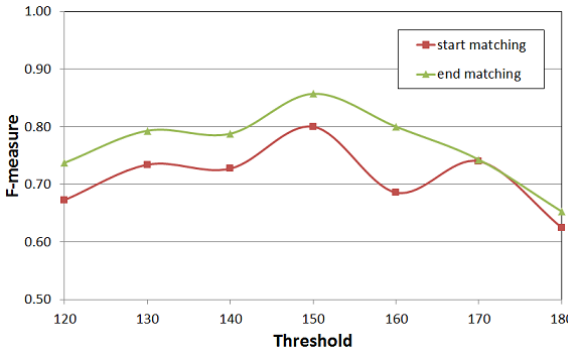

図 6　分割精度（F 値）

Figure 6　Segmentation accuracy (F-measure)

表 1,2 と図 6 より，閾値が 150 のとき，システムによる分割結果と人手による分割結果が最も一致し，開始タイミングと終了タイミングにおける F 値は，それぞれ 0.80 と 0.86 であった．

しかし，我々が講義を分割するために利用したジェスチャ動作による境界は, 文節の境界と完全に一致しておらず，実験において認識されたジェスチャ単位は，ほとんど文節よりも短く，部分的な一致にとどまっている．

6.　おわりに

本稿では，Web ブラウザ上で 3D 講師アバタを表示するインタラクティブな講義システムの提案を行った．また，そのシステムで利用するモーションデータを，モーションキャプチャ装置で, 講義を撮影することにより作成し, Web ブラウザ上で，モーションデータに基づく動作を行う 3D 講師アバタを表示するソフトウェアを開発し，ウェブブラウザ上で高品質な 3D モデルを利用して講義アニメーショ

ンを行うことが可能なことを確認した．

さらに，講義が利用するモーションのデータベース作成において，講師の発言に基づいてモーション分割をするために，ジェスチャ単位での分割を行った．ジェスチャ単位による分割結果は，人手による分割結果と非常に高い精度で一致することを示した．

しかし，ジェスチャ単位は文節よりも短く，部分的な一致となった．

そこで，我々はジェスチャ動作を利用した分割手法は，文節単位で講義を分割する際の必要な要素の一つだと考え，音量や音調，アクセントなどの情報を含んでいる音声データを利用した分割手法の両手法を併用することで，より効果的な分割を行うことが可能であると考える．

そして今後，本論文で紹介した技術をもとに，講義のモーションデータを自動的に蓄積していくだけでなく，人手でも講義の細かいモーションデータベースを作り，e-learning を行う学生のモチベーション向上にとって，適切な手法を提案し，その効果の測定を行う予定である．さらに，教育エージェントを利用したシステムや知的教育システムの研究分野において，エージェントの動作やエージェントを利用したシステムに関して研究を加速的に発展させるうえで，実際の講義の 3 次元モーションデータを蓄積していくことはとても重要となる．また，CG の分野において，人間らしい自然な動作を正確に定義し，それに基づいて動作を生成することは，膨大な時間を費やす作業である．そのため，我々が蓄積したモーションのデータベースを公開し，利用するといった方法の検討も行っている．

参考文献

1) WL. Johnson, JW. Rickel and JC. Lester: Animated pedagogical agents: Face-to-face interaction in interactive learning environments, International Journal of Articial intelligence in education 11.1, pp.47-78 (2000).

2) A. Ogan, S. Finkelstein, E. Mayeld, C. D'Adamo, N. Mat-suda and J. Cassell: Oh dear stacy!: social interaction, elaboration, and learning with teachable agents, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, pp.39-48 (2012).

3) D. Hasegawa, Y. Ugurlu and H. Sakuta: A Human-like Embodied Agent Learning Tour Guide for E-learning Systems, In Proceedings of 2014 IEEE Global Engineering Education Conference (EDUCON) Istanbul, Turkey, pp.50-53 (2014).

4) A.L. Baylor and K. Soyoung: Designing nonverbal communication for pedagogical agents: When less is more, Computers in Human Behavior 25.2, pp.450-457 (2009).

5) Biswas, G., Leelawong, K., Schwartz, D., Vye, N., & The Teachable Agents Group at Vanderbilt: Learning by teaching: A new agent paradigm for educational software, Applied Artificial Intelligence, Vol. 19, No. 3-4, pp.363-392 (2005).

6) 林勇吾, 小川均: Pedagogical Conversational Agent を用いた協同学習の促進 : 感情表出に着目した検討, 電子情報通信学会論文誌. D, 情報・システム J96-D, No. 1, pp.70-80 (2013).

7) Bikash Gurung, 新藤義昭, 松田洋: 3D-CG アニメーションを用いた対話型 e-Learning システムの開発と教材制作技法に関する研究, 情報科学技術フォーラム講演論文集, Vol. 8, No.3, pp.575-578 (2009).

8) McNeill, D. & S.Duncan, Growth points in thinking-for-speaking. In McNeill, D.(Ed.), Language and Gesture. pp.141-161, Cambridge: Cambridge University Press (2000).

9) McNeill, D., Cathments and contexts: Non-modular factors in speech and gesture production. In McNeill, D.(Ed.), ibid., pp.312-328 (2000).

10) D. McNeil: Hand and mind: What gestures reveal about thought, University of Chicago Press (1992).