情報システム 38－3
（1992. 3. 16）

音響認識による動画制御システム

川口昭良　杉本智香　白倉剛　千種康民　伊吹公夫

東京工科大学

視覚や聴覚など人間の知覚機構と関連したシステムを効率よく実現するには，人間の認識・行動を調査し，その機構に準じた方式を採ることが考えられる。筆者らは「音響認識による動画制御システム」を例にとったシステムで人間の認識機構の実験調査を行い，人間は「知覚対象を単一メディア環境だけでなく，マルチメディア環境としてで捉えていること」や，「認知機構が行動を判断する意図とも関連すること」などを確認した。このような人間の認識機構の性質を利用して，効率的な動画システムや，効率のよい認識システムが実現できた。そして，より複雑なシステムの構築にも発展できるようにモデル化したので，適用結果とともに報告する。

AUDIO CONTROLLED ANIMATION SYSTEMS

Akira Kawaguchi　Chika Sugimoto　Go Shirakura
Yasutami Chigusa　Kimio Ibuki

Tokyo Engineering University
1404-1 Katakura, Hachioji-shi, Tokyo 192, Japan

Investigation and survey of human cognitive and behavioral mechanism would be useful for efficient construction of an artifficial cognitive control system. We have found interesting answers, which suggested ' mutual interaction between multimedia perceptions ' and ' cognitive feedback from intention to act ', through thorough analysis of the audio controlled animation prototype system exhibition, and obtained a useful model for general applications. The cognitive experiments and the obtained model with application results, are discussed in this paper.

１．はじめに

視覚や聴覚など人間の知覚機構と関連したシステムを効率よく実現するには、人間の認識・行動を調査し、その機構に準じた方式を採ることが考えられる。筆者らは、「音響認識による動画制御システム」を例にとったシステムで人間の認識機構の実験調査を行い、人間は「視覚対象を単一メディア環境だけでなく、マルチメディア環境として捉えていること」や、「認知機構が行動を判断する意図とも関連すること」などを確認した。

そして、このような人間の認識機構の性質を利用して、効率的な動画システムが構築できた。また逆に、この認識機構を分析してモデル化し、これを人工的な認識機構の構築に活用して、効率の良いシステムも実現できた。本システムで試みた手法は、より複雑なシステムの構築にも発展できる可能性があると思われるので、そのモデルと共に適用結果を報告する次第である。

２．システムモデル

人間の認識・行動を図１のように、知覚機構－認知機構－行動機構というモデルで捉える[1]。さらに、マルチメディアとしての知覚器官や行動と関連した意識などの相互作用を受けたり、情報理論にかなった母集団の統計的予測を無意識の内に行ったりして、効率の良い認識を行っていると言う仮説をたてる。その結果、図２のようなシステムモデルを描いてみた。人間の機構がこのようなモデル通りか否かは別としても、効率の良い人工システムが得られることは、応用例の試作実験から確認できた。

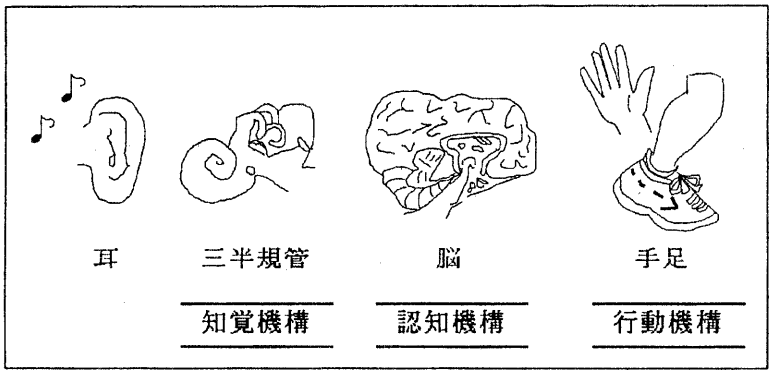


図１　認識行動モデル

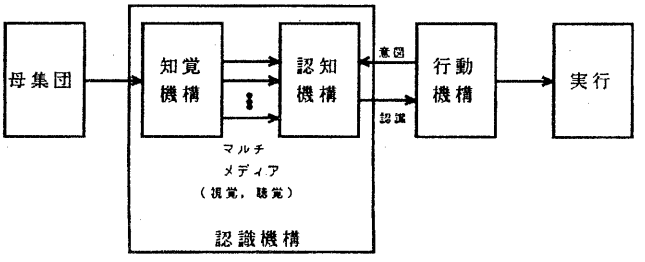


図２　システムモデル

3．実験システムの構築

「単一メディア環境」、「マルチメディア環境」、「行動との結合環境」の三種類の認識環境に対応して、表1に示すような実験システムを用意した。

表1　実験システム

認識の環境	実験システム
単一メディア	棒の回転
マルチメディア	人形のダンス
行動との結合	もぐら叩きゲーム

4．動画の認知科学的実験

棒の回転の動画（図3）をセル数を変えて、アンケートにより自然性を調査した結果を表2に示す。

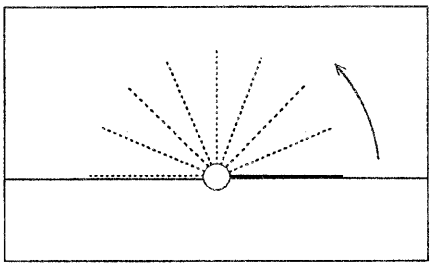


図3　　棒の回転

表2　アンケート結果

セルの枚数	1コマの回転角度	回転しているように見える	回転しているように見えない
9	22.5°	84%	16%
6	36.0°	67%	33%
5	45.0°	57%	43%
4	60.0°	19%	81%
3	90.0°	9%	91%

単一メディアにおける角度と自然性の関係をグラフにすると、図4のようになるが、人形のダンスの場合には、見学者の多くの意見より、図の破線の位置でも自然性が高く評価されることが確認された。これは音楽を聞くことによって、動きを予測するためと考えられる。視覚聴覚を複合したマルチメディア環境と視覚だけの単一メディアとの相違点である。

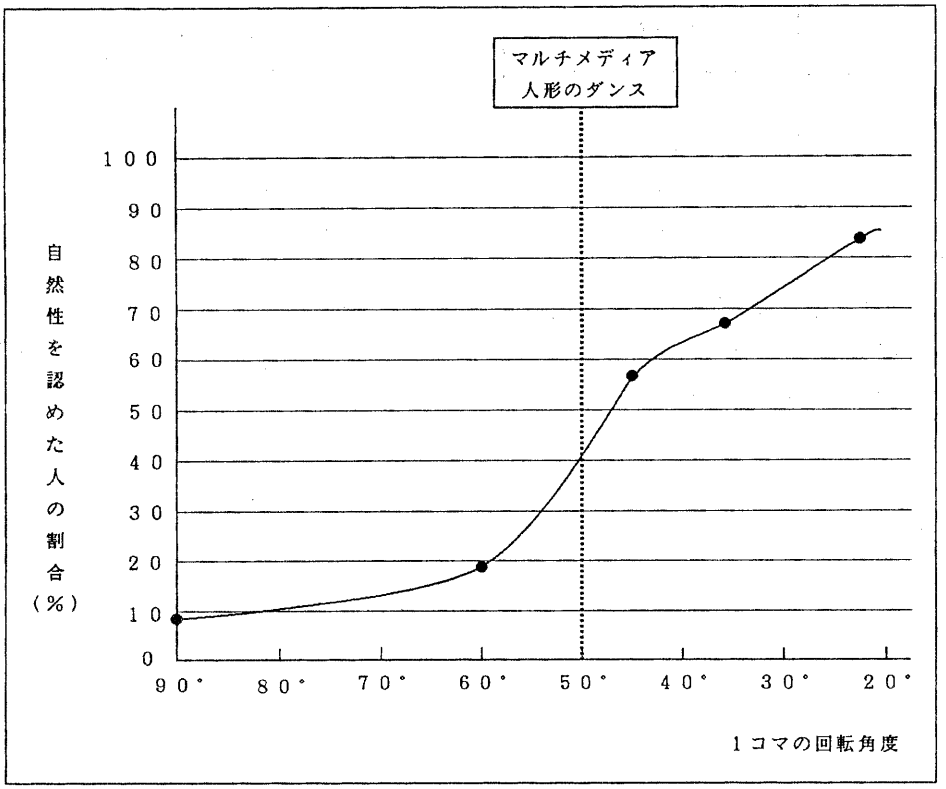


図4　角度と自然性

このことは、行動を伴う場合でも同様の現象がある。出てくるもぐらの位置に応じた音のキーを叩くという「もぐら叩きゲーム」では、動画面に神経を集中し、動きを予測している為、セル数が少なくても満足な成果がえられた。これも画面の予測をする結果と考えられる。

5．人間の認識機構

人間は音を聞いたり、ものを見るときに、背景となる雑音のある環境の中で、背景を無視し、着目したものだけに照準を合わせて認知する。これは対象の予測に基づいて対象以外をマスクすることにより、精度を向上していると解釈できる。今述べた動画の認知科学的実験でも、置かれた環境によって自然性が異なるのは、認知する対象の予測の度合いが異なるものと解釈できる。

このような予測は、「知覚機構ではなく脳で行っている」と考えた方が自然である。そこで、人間の認識機構を図２のモデルを想定して説明することができる。

6．音響認識への応用

音響認識を行うには、人間の三半規官に準じて、音圧の関数を時間領域から周波数領域に変換するのが自然である。周波数変換には、通常フーリエ変換を採用するが、時間軸を有限に制限するとウィンドウの切り方に依存した誤差があり、情報理論的な考慮をしたＭＥＭ法は、これを緩和する解析的な方法である。但し、ＭＥＭ法では実時間での制御には適さない。そこでフーリエ変換を用いても認識対象を限定することにより、十分な解析を行うことができる。また、入力信号の統計的性質を効率よく利用するには、知覚器官に対応する解析的な処理だけではなく、頭脳に対応する論理的な情報処理を付加したモデルも考えられる。図５は、「音響認識による動画制御システム」を具体的にしたものである。これは図２のモデルを音響認識に適用したものである。

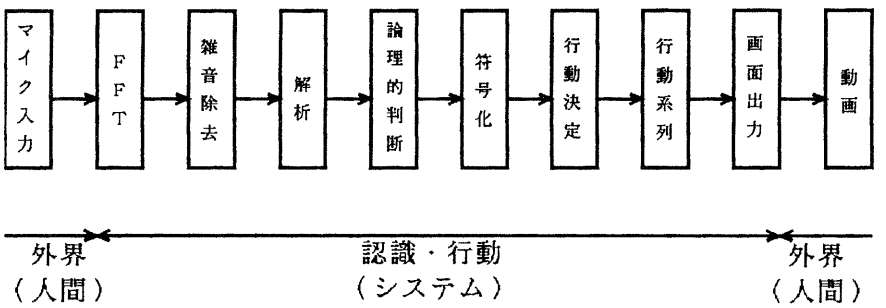


図５　音響認識による動画制御システム

一般楽器に対応しては音色を判断する機構、音声に関しては文脈の判断などのより高度な論理的処理が必要となろうが、今回の試作では原理モデルの確認のため、もっとも簡単な「雑音のある環境下で、フルートの単音、和音までが認識できるもの」を対象に選んだ。そして、制約条件を利用した論理的判断と高速フーリエ変換とを組み合わせて、小型機でも実時間処理ができる効率の良い方法を適用し実用性を満足することを確かめた。その原理と応用の仕方を実例で説明する。

７．音響認識部

７．１　音響認識部の構成

図６に音響認識部のシステム構成を示す。

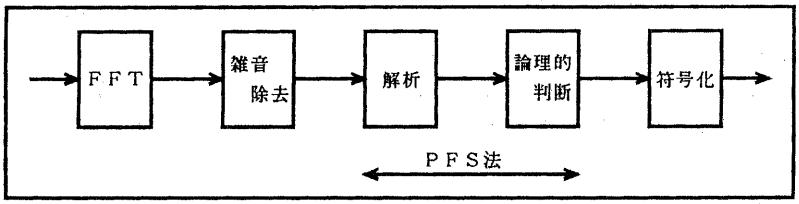


図６　認識部のシステム構成

７．２　雑音除去

ＦＦＴによって得られたスペクトルの総和f_{sum}とスペクトルの最大値f_{max}の関係が、

$$f_{max} > \alpha f_{sum}$$

を満足する時f_{max}は、有効な信号とみなして次の処理に進む。但し、実験結果により今回は$\alpha = 0.05$と定めた。

７．３　ＰＦＳ法

音源の統計的性質を生かした予測により、論理的判断をするものをＰＦＳ法（ピッチ周波数スペクトル法）と呼ぶことにし、この原理を図７に示す。

フルートの場合には高調波を無視し基本波だけで判断できるので、音の認識は、次の処理によって行う。

サンプリング周波数ｆｓで正規化した周波数軸上の認識を意図した音の周波数をｆｒとする。この両隣りのスペクトルｆ１，ｆ２のスペクトル値をそれぞれＶ１，Ｖ２、ｆｒからの周波数軸上の距離をそれぞれＲ１，Ｒ２とする。

これからｆｒのスペクトル値を次の式で求める。

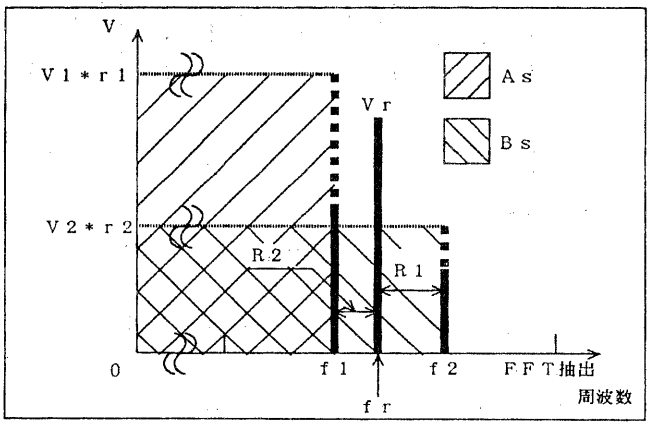


図７　ＰＦＳ法の原理

Ｒ１≧Ｒ２　なら、

$$Vr = \{\beta f1 V1 (R1+1) + f2 V2 (R2+1)\} / 2 fr$$

Ｒ１＜Ｒ２　なら、

$$Vr = \{f1 V1 (R1+1) + \beta f2 V2 (R2+1)\} / 2 fr$$

但し、　　ｆｒ≦　２５　　なら、　β＝５５
　　２５＜ｆｒ≦　５０　　なら、　β＝７９
　　５０＜ｆｒ　　なら、　β＝９０

Ｖｒの算出は、Ｖ１，Ｖ２に重みを付加し、加重平均したものである。加重の仕方は、認識率が、高くなるように実験的に求めた。

このＰＦＳ法の、大きな特徴は、四則演算しか用いてない点である。このことは、実時間性に大きく貢献している。

７．４　符号化

いま説明した判定結果を符号化し、図８に示す動画部への引継情報を得る。

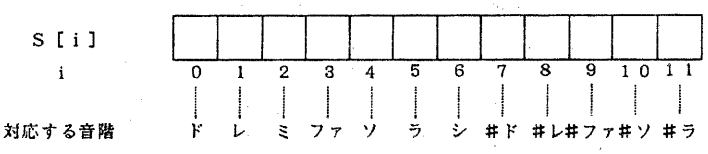
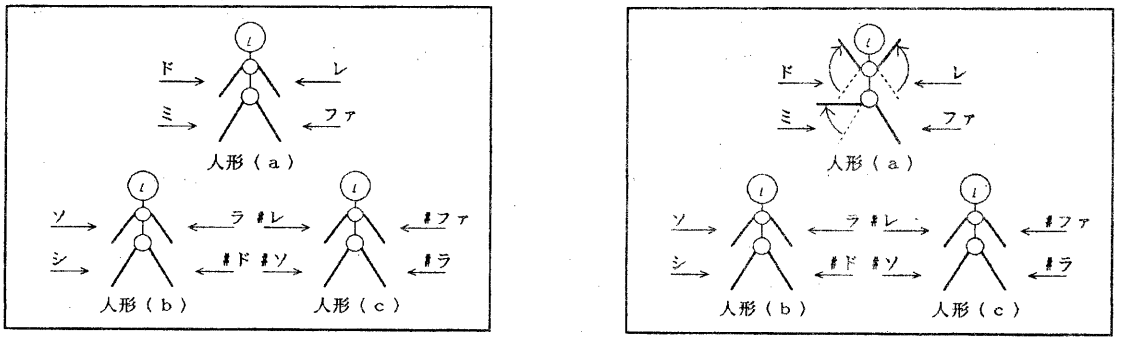


図８　引継情報

８．動画制御部

８．１　動画制御部の構成

動画制御部では、図５の行動の決定，行動系列の作成，及び画面出力までを取り扱う。図８の引継情報を受けて、必要な動画を作成する。人形のダンスを例にとり説明する。



（ａ）行動前　　（ｂ）行動後（ﾄﾞ･ﾚ･ﾐ和音入力時）

図９　人形のダンス

人形のダンスは音の種類により図９のような動作をする。このために、画面を図１０のようなセルに分割し、一つの引継情報に対して、定まった時間的順序でセルを表示する。この際、セル数が少ないほど、アニメーションの作成費が節減でき、また、記憶容量や制御処理量が少なくてすみ、家庭用の小型機でも容易に実現できることになる。効率的な動画の実現には認知科学的性質を配慮したセル割りの設計が大切である。

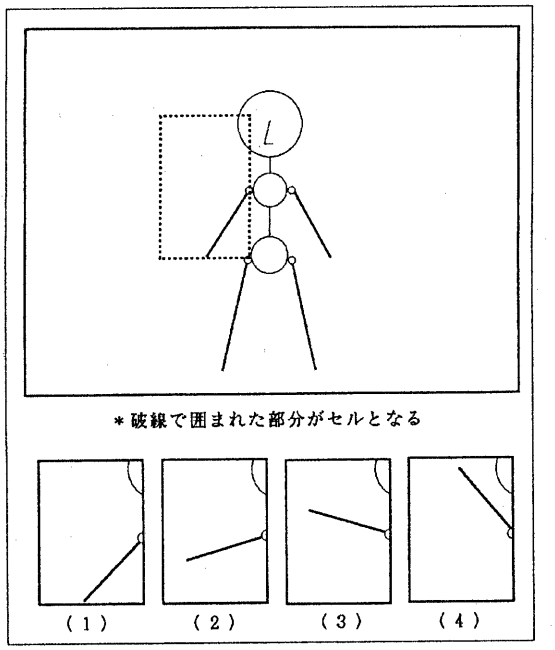


図１０　セル分割

8．2　行動系列の認知科学的設計

単純なセル割りは、表２の結果に基づいて設計できる。マルチメディア環境や行動と結合した環境ではさらに少ないセル数で充分である。このように、環境と対象との関係を配慮することによって効率的な画像を作ることができる。

９．実時間性

以上述べた認知科学的性質を活用して効率のよいシステムの実現を試みたが、その効果を、実時間性という立場から、定量的に評価する。

今回試作した人形のダンスにおける音響入力から動画出力までの１回の処理時間は、０．４６２秒であり、実時間性を満足した。内訳は図１１の通りである。

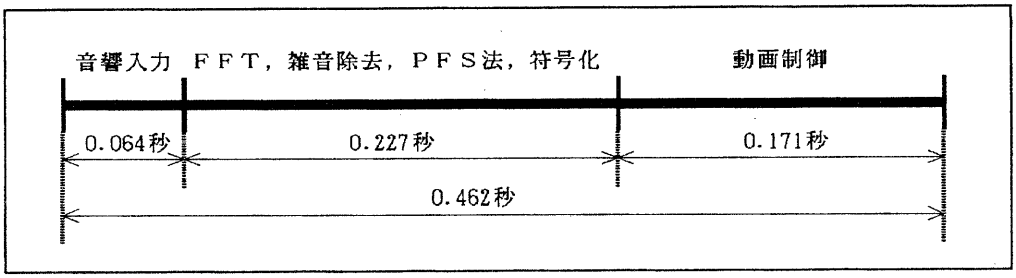


図１１　処理時間

１０．おわりに

本報告は視覚分野について、環境と対象の認知科学的整合による時間軸方向の効率的な（記憶量や伝送量や処理量からみた）出力法を報告した。多階調画像と二値画像との相互変換という空間軸方向の問題に関しては、基礎的な調査実験を別途報告している。今後、文字、風景、人物、設計図など、対象の性質に即した手法の論理的選択や学習機能の取り入れなど、本報告の手法と総合した発展が期待される。また、聴覚分野でも、より高度な論理的判断を取り入れ、今回対象にしなかった応用領域への発展も考えられる。

謝辞

モデルの表現に関して心理学的立場から御教示いただいた東京工科大学奥正広講師、またアンケート調査を担当して頂いた近松康子女史、及びアンケートに御協力頂いた多数の方々に感謝する。

参考文献

[1] D.Norman : The Psychology of Everyday Things, Basic Books, 1988.