複数の異なるモダリティ情報の統合に基づく 移動ロボットの行動選択

正 員 陳 彬* 非会員 金子 正秀*

Behavior Selection of Mobile Robot Based on Integration of Multimodal Information

Bin Chen*, Member, Masahide Kaneko*, Non-member

Recently, biologically inspired robots have been developed to acquire the capacity for directing visual attention to salient stimulus generated from the audiovisual environment. On purpose to realize this behavior, a general method is to calculate saliency maps to represent how much the external information attracts the robot's visual attention, where the audiovisual information and robot's motion status should be involved. In this paper, we represent a visual attention model where three modalities, that is, audio information, visual information and robot's motor status are considered, while the previous researches have not considered all of them. Firstly, we introduce a 2-D density map, on which the value denotes how much the robot pays attention to each spatial location. Then we model the attention density using a Bayesian network where the robot's motion statuses are involved. Secondly, the information from both of audio and visual modalities is integrated with the attention density map in integrate-fire neurons. The robot can direct its attention to the locations where the integrate-fire neurons are fired. Finally, the visual attention model is applied to make the robot select the visual information from the environment, and react to the content selected. Experimental results show that it is possible for robots to acquire the visual information related to their behaviors by using the attention model considering motion statuses. The robot can select its behaviors to adapt to the dynamic environment as well as to switch to another task according to the recognition results of visual attention.

キーワード：視聴覚情報，ベイジアンネットワーク，行動選択，動作状態，マルチモーダル統合，全方位移動ロボット，視覚的注意

Keywords: Audiovisual information, Bayesian network, Behavior selection, Motion status, Multimodal integration, Omni-directional mobile robot, Visual attention

1. はじめに

近年，ユーザと人間共存型ロボットとの間のフレンドリーなインタフェースを実現するために，知能ロボットに顔，頭，アーム，音声機能などを持たせ，表情，ジェスチャーなどを自動的に生成する研究が活発に行われている⁽¹⁾⁽²⁾⁽⁵⁾。特に，ロボットの頭部や眼球の動きがユーザに与える印象は，顔表情と共に重要であるため，視覚的注意を工学的にモデリングすることによって，ロボットの頭部や視線の動きを制御する方法に関心が高まっている⁽⁶⁾⁽⁸⁾⁽¹¹⁾。知能ロボットに視覚的注意を導入することには，以下の2つの利点がある。

ある。①知能ロボットが，画像内に含まれている視覚的に顕著なオブジェクト，あるいはロボット自身の行動状態やタスクと緊密している視覚情報を，効率的に素早く抽出し，それに応じた行動選択を行うことができる⁽⁹⁾⁽¹⁰⁾⁽¹²⁾。②ロボットが，眼球の動き，顔の姿勢，ジェスチャーなどを用いてシーンに対する注意を表すことによって，ユーザがロボットの注意対象を共有することができる。これによって，ユーザとロボットとの間での相互コミュニケーションの柔軟性を向上させることができる⁽⁷⁾。

1・1 関連研究 人間の視覚的注意のメカニズムを解明する研究，並びに，ロボットへの実装は従来から行われている^{(8)~(12)}。文献(8)(11)では，ロボットが特定の色，画像内の動きなどに対して注意を選択し，頭部や眼球の動きを制御する。文献(9)(10)では，屋外画像を主な対象として，画像内の画素を中心として，周辺領域と中心領域との画像特徴の差を計算し，特徴顕著度を表すマップを作成

* 電気通信大学 大学院電気通信学研究科 電子工学専攻
〒182-8585 東京都調布市調布ヶ丘1-5-1
Department of Electronic Engineering, The University of
Electro-Communications
1-5-1, Chofugaoka, Chofu-shi, Tokyo 182-8585, Japan

する。これによって、人間に似たような注意点選択を行うことができる。しかし、知能ロボットが環境からの聴覚的刺激に対する反応を起こすためには、視聴覚情報を統合した注意モデルを構築する必要がある。

人間は視聴覚情報を統合して、モダリティ間での相互作用によって視覚的注意点を選択する⁽¹⁸⁾。このような相互作用を確率の方法などによって工学的にモデリングする研究があるが、実空間内のロボットに実装して応用することは困難である^{(15)~(17)}。これは、一般に視覚空間と聴覚空間は異なっており、両空間の間での相互マッピングを行う必要があるためである⁽¹⁹⁾。すなわち、視聴覚情報環境下でロボットに注意選択をさせるためには、位置関係が互いに一致した視聴覚空間で画像、音声信号の統合処理を行うことが不可欠である。

注意選択においては、視聴覚情報の他、行動状態を含めた3モダリティの情報を統合して、行動に緊密した視覚情報を抽出することが行われている^{(20)~(22)}。これらの研究では、移動ロボットの進む方向を入力とした2つのポインターニューロン (Pointer Neuron) からの出力を利用して、注意分布密度をモデリングする。注意分布密度マップと画像顕著度マップとの結合によって、ロボットに移動方向上の領域を優先的に選択させる。従って、ロボットは、ナビゲーションという行動に関連した画像情報を、それ以外の情報に比べて優先的に選択することができる。しかし、ロボットの外界環境は3次元であり、奥行き情報を考えないで移動方向上の領域に濃い注意密度を分布させることは必ずしも適切ではない。また、文献(22)で用いられているロボットの行動環境は、白い背景や積み木などのオブジェクトから構成されており、比較的簡単である。

1.2 本論文での方法 本論文では、移動ロボットの行動状態、画像特徴情報、シーンの奥行き情報、音源位置センシングの結果を統合して、行動に関連する実空間内の画像領域を優先的に選択させる方法を提案する。まず、ロボットの行動状態、シーンの3次元情報をノードとしてベイジアンネットワークを構築する。空間内で移動する被験者における注意点の選択結果をサンプルとして、ベイジアンネットワークを用いて分布様式を学習する。次に、ステレオ視とマイクロホンアレーを併用した音源位置推定方法を用いて、画像座標系と一致した空間で音源を定位する⁽¹⁴⁾。これによって、従来の研究において必要であった視聴覚空間の相互マッピングは不要になる。また、視聴覚情報と注意密度分布マップを結合し、積分発火ニューロンで積分を行う。発火したニューロンにおける画像領域をロボットの注意領域とし、それに対してロボットが反応する。このように、3つのモダリティからの情報を統合することによって、視聴覚的に顕著な特徴だけでなく、行動に関連した画像領域に対しても、ロボットが注意を向けることができる。最後に、選択された注意領域に対する画像認識を行い、ロボットは実空間内にいるユーザや障害物によって行動を自動的に調整する。これらの実験を通じて、本論文での提案

図1 ハードウェア構成 (a) 全方移動ロボット, (b) 正面図, (c) 側面図

Fig. 1. Hardware configuration. (a) Frontal view of the omni-directional mobile robot. (b) Frontal view and (c) side view of stereo camera and microphone array configuration.

方法を用いることによって、ロボットがマルチモダリティ情報源の間で、行動に関連した注意選択を行うことができることを明らかにする。

まず、2. で本論文で使用するハードウェア構成について説明する。3. では、ベイジアンネットワークを利用して、ロボットの行動状態とシーンの3次元構造から視線選択への影響をモデリングする方法について述べる。4. では、シーンのカラー画像、音声情報、行動状態に基づいて、視野内の各空間領域がロボットの視覚的注意を引く度合を計算し、ロボットの視覚的注意先を選択する方法について述べる。5. では実験を行い、視聴覚情報環境下での視覚的注意の自動選択及び注意対象の認識に基づく行動選択が重要であることを明らかにする。6. で本論文をまとめる。

2. ハードウェア構成

全方移動ロボットに人間とのコミュニケーションに必要な視聴覚センサーを装着した状態でのハードウェア構成を図1に示す。ロボットの頭部には立体視可能な3眼カメラが取り付けられており、環境の奥行き距離を測定することができる。ロボット本体の正面に5チャンネルのマイクロホンアレー (図1に印で示してある) が実装されており、画像情報と結合して、カメラによる可視音源の位置情報をセンシングすることが可能である。ステレオカメラ座標系における各マイクロホンアレーの位置座標は図1(b)(c)に示すように事前に測定してある。また、本体の中央部にマイクロホンを接続したアンプとA/D変換器を積み重ねている。3眼カメラ、マイクロホンアレー、駆動系はそれぞれ異なるPCと接続されている。複数のPCはネットワークを介して情報交換を行う。

3. 空間領域におけるロボットの注意密度の分布

人間における視覚的注意の選択は、視聴覚情報により影響されているだけでなく、行動状態による影響もある。たとえば、前へ進む人間は自分の移動経路上に高い注意を

払っているが、それ以外の領域に配分する注意は少ない。このような注意配分は、画像やオブジェクト特徴によるものではなく、人間自身の行動状態からの影響を受けたためである。そこで、ロボットが実空間内で行動する際、行動状態に応じた空間内各箇所への注意配分をモデリングし、視聴覚情報と統合することによって、行動に関連した注意点を選択することができるようにさせる。

3・1 知能ロボットにおける 3 次元空間領域への注意配分の重要性 3 次元空間への注意配分は各領域によって異なる。注意配分の多少を数値化したものが「注意密度」である。一般的に、観測者の近くに配分される注意の密度は濃く、逆に遠くに配分される注意の密度は薄いと考えられている⁽¹³⁾。また、観測者が静止した状態では、配分された注意は全体的に薄く、他方、動いている状態では濃くなる。さらに、行動の状態、例えば移動速度や方向などによって、3 次元空間における注意密度の分布は異なる⁽¹³⁾。

知能ロボットは、行動状態やタスクなどといったトップダウンの情報、及び視聴覚的刺激といったボトムアップの情報を統合することによって注意先を選択する。どこに注意を向けているのかに応じて外界の情報を効率的に取得し、処理することができる。例えば、選択された注意先に基づいて、知能ロボットが 3 次元空間からの画像情報を異なる解像度で取得することは、情報処理の効率化のための重要な手段であり、アクティブビジョンの分野においても関心が高まっている⁽⁴⁾。

そこで、以下では、ロボットの行動状態、及び、シーンの 3 次元情報に基づいて、ベイジアンネットワークを用いて空間へのロボットの注意配分密度をモデリングする方法について述べる。この方法によって、従来の研究では実現されていなかった、3 次元環境と行動状態という要素が様々な組合わさった状況下で、ロボットの環境への注意配分を数値化することができる。

3・2 ベイジアンネットワークによる注意密度のモデリング ベイジアンネットワークは、事象をノード、各事象間の因果関係をリンクにより記述した有向グラフである。学習サンプルを用いて、各ノードの条件付き確率表 (CPT: Conditional Probability Tables) を学習する。各ノードの値を観測した場合、ネットワークの条件付き確率表に基づいて、目的となるノードでの値の確率分布を推定することができる。ロボットが実空間内で走行や回転をする際、空間内の各箇所への注意の配分密度は、本体とその箇所との距離、相対角度と言った外的要因、及び、ロボット自身の移動・回転の速度、方向と言った内的要因に深く係わっている。例えば、ロボットの移動方向、速度に応じて、移動経路上にある、本体との距離が近い物体に対する注意密度は、それ以外の空間より比較的濃い。一方、これに対応する人間の注意配分システムは、上述した内外的要因の条件下で行動やタスクに緊密した注意領域を選択する最も優れた例である。このため、ロボットの環境への注意分布は、人間における注意点の選択結果から学習させることができる。

図 2 注意分布密度をモデリングするベイジアンネットワーク

Fig. 2. Bayesian network of modeling distribution of visual attention density.

本論文では、ベイジアンネットワークを利用して、異なる行動状態とシーンの 3 次元構造を内外的要因として、人間における注意領域の選択結果を記録して、統計的手法により各要因が組合わさった条件下での注意密度の分布状態を調べる。

3・3 グラフの構造とパラメータ 本論文では、上述した内外的要因を事象とし、各事象間の因果リンクにより図 2 に示すベイジアンネットワークを構築する。人間の注意選択の結果を学習例として、ベイジアンネットワークを学習させる。この結果、ある行動状態下でのロボットが実空間中のある箇所へ配分する注意密度は、その領域とロボットとの相対位置関係、学習したベイジアンネットワークのパラメータにより計算される。

図 2 のベイジアンネットワークに含まれている各ノードに対する説明を表 1 に示す。ただし、“x”はその欄に対応する値が取られていないことを示す。このグラフについて以下に詳しく説明する。

(1) ノード “Direction” は、ロボット本体の直進角度を示す。ここでは、カメラの光軸を中心とした $[-75^\circ, 75^\circ]$ だけを考慮する。さらに、移動角度を表 1 に示すように 1~5 の整数値に離散化する。

(2) ノード “Azimuth” は、実空間中のあるオブジェクトとカメラ光軸との方位角である。シーンの距離画像上のある画素 $p(i, j)$ に対応する実空間中の座標を $P(x, y, z)$ とすれば、 $\text{Azimuth} = \arctan(x/z)$ である。本論文で用いたカメラの視野角は 50° であり、方位角を表 1 に示すように 1~5 の整数値に離散化する。

(3) ノード “P-Obstacle” は、潜在的障害物を表す、観測不能なノードである。このノードが真である場合、この空間箇所とロボットとの相対距離、及び、ロボット本体の移動速度により、実際の障害物 (Obstacle=1) になる可能性がある。

(4) ロボット本体の移動・回転速度の範囲はそれぞれ、 $[1, 10]$, $[-5, 5]$ である。視差 “Disparity” については、 $[0, 60]$

表 1 ベイジアンネットワークの各ノードの説明

Table 1. Description of each node in the Bayesian network.

Number	Node Name	Description	Value				
			1	2	3	4	5
1	Direction	Moving Direction of Robot	$[-75^\circ - 45^\circ]$	$[-45^\circ - 15^\circ]$	$[-15^\circ 15^\circ]$	$(15^\circ 45^\circ)$	$(45^\circ 75^\circ)$
2	Azimuth	Angle between Point P and Optical Axis of Camera	$[-25^\circ - 15^\circ]$	$[-15^\circ - 7^\circ]$	$[-7^\circ 7^\circ]$	$(7^\circ 15^\circ)$	$(15^\circ 25^\circ)$
3	P-Obstacle	Possible Obstacle	True	False	×	×	×
4	Disparity	Disparity of Point P	$[0 10]$	$[10 18]$	$[18 25]$	$(18 60]$	×
5	Speed	Moving Speed of Robot	0	$[1 5]$	$(5 10]$	×	×
6	Obstacle	Obstacle	True	False	×	×	×
7	Turn	Rotation of Robot	$[-5 0]$	0	$(0 5]$	×	×
8	Gaze	Visual Attention	True	False	×	×	×

の区間から $[1 5]$ にマッピングする。

(5) ノード “Gaze” は、ロボットが視覚的注意をこの箇所に向けたかどうかを示す。ベイジアンネットワークにより得られた $Pr(i, j)$ ($Gaze = 1$) は、この箇所を注意点として選択した場合の尤度を表す。本論文では、この尤度を用いて注意密度を表す。以下では、 $Pr(i, j)$ ($Gaze = 1$) を単に $Pr(i, j)$ で記す。

表 1 から、本体の回転、直進速度、直進方向に関する 3 つのノードの値を組合せることにより、計 45 通りのロボットの基本行動パターンがあることが分かる。このように多くのパラメータを決める問題において、学習という方法が有用であるが、教師データの収集が重要となる。教師データを得ることが困難である場合には、ロボットと同様な条件下で人間の選択結果を計測して、学習データとすることによって、システムの構築を比較的容易にすることができると考えられる。このような観点から、本論文では、ロボットをこれらのパターンに従って行動させ、カメラで取得した時系列の距離画像を記録する。記録した距離画像に対して、ユーザは、ロボットの移動に障害となる可能性がある、すなわち、ロボットにとって最も注意を要する領域を 1 つ選択して記録する。このような時系列画像を複数回再生し、データ収集を繰り返して行う。収集した計 50000 個の学習データを用いて EM アルゴリズムにより CPT を学習させる。

3.4 移動中のロボットにおける注意密度の分布 本論文では、シーンの距離画像を 80×60 の解像度に変換する。変換後の距離画像上の各画素 $p(i, j)$ に対して、ベイジアンネットワークを適用して $Pr(i, j)$ を求める。ここで、移動中のロボットにおける注意密度の分布マップ $Pr(i, j)$ の計算結果を示す。

図 3 は、第 106 フレームで得られた注意密度の分布結果を示している。第 106 フレームでは、ロボットの本体が画面の左奥へ向かって移動する。すなわち、行動状態は $Direction = -30^\circ$ 、移動速度 $Speed = 3$ 、回転 $Turn = 0$ である。図 3(a)(b) の各々は、シーンの画像と視差マップを示す。各画素に対して、その画素における視差値、行動状態をベイジアンネットワークに入力して $Pr(i, j)$ の値を

図 3 注意分布密度の計算結果 (a) シーン画像、(b) 距離画像、(c) 注意密度マップ

Fig. 3. The distribution of attention density acquired by Bayesian network. (a) Image captured by camera. (b) Depth map. (c) Distribution map of attention.

計算する。計算した 2 次元マップ Pr を図 3(c) に示す。第 106 フレームの場合、ロボットの前方に立っている人物、及び、人物のそばにある机がロボットの移動を妨げる可能性があるため、この領域に注意密度の分布が高く、図 3(c) に示す注意密度マップにおいて明るい領域により示されている。一方、この方向の逆側 (画面の右側) では注意密度の分布が低くなっていることが分かる。

4. 視覚的注意点選択のための視聴覚情報と行動状態との統合

ロボット自身の動き状態とシーンの 3 次元構造が注意点の選択を左右する要因である。シーンに関する画像や音響的变化を捉えるために注意点を選択する際には、シーンの画像特徴、音源位置なども考慮する必要がある。ここでは、移動中のロボットが、環境からの視聴覚情報の刺激に基づいて空間内にあるオブジェクトに注意を向けるための、視線選択方法を提案する。

4.1 視覚情報の処理 単に画像情報を利用する場合、人間の視覚における選択的注意機能をモデル化するためには、画像内の色特徴、エッジの方向性、対称や円形といった画像内のオブジェクトの構造的特性などの特徴マップを線形的に結合する方法がある⁽¹⁰⁾。この方法では入力画像の多重スケール画像から特徴を抽出し、特徴マップを生成する。各特徴マップ上の任意の領域を中心領域としてその周辺領域との差を計算して 2 次元のマップを生成する。各マップの和を計算し、マップ $I_v(i, j)$ とする。高い $I_v(i, j)$

を持つ領域は周辺領域と比べて顕著な特徴を有するため、視覚的注意が最も引かれる領域である。本論文では、画像の色特徴(肌色情報, 赤, 黄, 緑), Gabor フィルターによるエッジの方位性特徴 ($0^\circ, 45^\circ, 90^\circ, 135^\circ$), 動き情報を画像特徴として用い, 上述のマップ $I_v(i, j)$ を計算する。さらに, 4・3 でマルチモーダル情報間の統合を行うために, マップ $I_v(i, j)$ 上の値を $[0, 1]$ の区間に正規化する。

4・2 音源センシング 画像座標系と一致した空間で音源位置を推定することができれば, 視聴覚両空間の間で相互マッピングを行う必要がなくなる。すなわち, 画面から音源位置を直接検出することによって, その後の視聴覚情報統合の複雑さを低減することができる。

4・2・1 画面からの音源位置検出の方法 本論文では, 図 4 に示すように, 視聴覚情報の統合により音源位置を画像から検出する方法を利用し, 音源位置センシングを行う⁽¹⁴⁾。図 4 の“MICs”では, 多チャンネル音声信号に対して白色化相互相関を計算した結果を, シーンの距離画像によって 2 次元のマップに変換する。このマップは音源位置の尤度を表しており, これによって音源位置を推定することが可能であるが, 白色化相互相関は環境内の雑音や反響などの影響を受けやすいため, 音源位置を誤って推定する場合がある。そこで, このマップを音源の視覚的情報と統合することにより, 音源位置をさらに安定して推定することを図る。図 4 中の“SVM”は, 音声信号の分類処理を表す。一般的に, 音声音の音源位置と画像内の肌色分布との相関が強く, 非音声音の音源位置と肌色分布との相関はそれほど強くないと考えられる。これは, 肌色領域に話者の顔が存在する可能性が高いためである。図 4 中の“GMM”は, 肌色情報に対する処理を示す。また, ユーザとのインタラクションにおいては, 環境内に音声音が多く存在するが, 非音声音信号もある。非音声音の場合, 音源と肌色との関連性が低いため, “MICs”で得られた音源位置を表す尤度マップを肌色尤度マップと結合する際, 音の種類によって結合重みを自動的に調整する必要がある。図 4 中の“SVM”は, サポートベクターマシーンを利用して音声信号を音声音/非音声音に分類する処理を示す。また, 同図の“Integration”は, “MICs”, “SVM”, “GMM”の 3 者からの出力情報を確率的に統合する処理を示す。このようにして, 音声音/非音声音の音源位置を画像から安定して推定できるようにする。ここで, “Integration”から出力した音源位置の尤度マップを $R(i, j)$ により示す。

図 5 は音源位置推定の結果を示している。図 5(a) は, 画面の奥側にいるユーザが本を読んでいるシーンを示す。(b) は上述した視聴覚情報の統合によって得られた音源位置に関する尤度マップ $R(i, j)$ である。このマップでは, 尤度値を明るさで表しており, 最大尤度値を検出して音源の位置を推定することができる。推定した音源位置を (a) のシーンの画像に示す。この実験から, 音源位置を画像平面から直接検出できることがわかる。また, ロボットにおける視覚空間, 聴覚空間が画像平面に統一され, 音声信号の刺

図 4 音源位置センシングの方法

Fig. 4. The method of sound source localization

図 5 音源位置センシングの結果 (a) シーン画像, (b) 音源位置を表す尤度マップ

Fig. 5. The result of sound source localization. (a) Image captured by camera. (b) Spatial likelihood map of sound source location.

激により, 視覚的注意対象を検出し, 視線を向けることが可能になる。

4・2・2 音源誘目度マップの生成 音声信号によってロボットの視覚的注意を選択する際, 音声信号のパワーや高次の意味情報などを考慮する必要がある。音声信号の意味的情報が視覚的注意の選択に与える影響を工学的にモデリングすることは困難であるため, 本論文では, 音声信号のパワーだけを利用する。また, 異なる空間位置で発せられた異なるパワーの音声信号がロボットの視覚的注意を引き起こす度合いを, 音源誘目度によって表す。

画面から検出した音源の位置を $p_s(i, j)$ とすると, ステレオ視により画素 $p_s(i, j)$ が対応している音源の 3 次元位置 $P_s(x, y, z)$ を求めることができる。そこで, 幾何学的関係から, 音源 $P_s(x, y, z)$ から発せられた音声信号の各マイクロホンペアへの到来時間差 τ を式 (1) によって計算することができる。ここで, m, n はマイクロホンの番号, τ_{mn} は両マイクロホンまでの音の到来時間差を示す。 P_m, P_n, F, c の各々は, ステレオ座標系における両マイクロホンの 3 次元座標, 音声信号のサンプリング周波数, 音速である。

音声信号の各フレームで音源位置に指向性を形成して目的音を強調する。このためには, 遅延和アレー (Delay-and-Sum: DS) を用いる。遅延和アレーでは, 各マイクロホンで受信した信号に遅延を付加した後, 総和を取ることで, 目的とする位置の音声信号を強調する。式 (2) で, s^c は強調した音声信号を示す。 N, n, s_n はマイクロホンの個数, 番号, 第 n 番目のマイクロホンの音声信号を示す。

$$\tau_{mn} = \frac{(|P_s - P_m| - |P_s - P_n|)F}{c} \dots\dots\dots (1)$$

$$s^c(t) = \frac{1}{N} \left(s_1(t) + \sum_{n=2}^N s_n(t + \tau_{1n}) \right) \cdots \cdots (2)$$

$$\lambda = \frac{E(t)}{E_{max}} \cdots \cdots (3)$$

$$I_a(i, j, t) = \lambda \frac{e^{-\frac{1}{2}(\mathbf{x} - \mu_s)^T \Sigma_s^{-1} (\mathbf{x} - \mu_s)}}{\sqrt{(2\pi)^d |\Sigma_s|}} \cdots \cdots (4)$$

各フレームの音信号 $s^c(t)$ に対して、パワー $E(t)$ を計算し、式 (4) に示す 3 次元のガウス関数により音源誘目度マップを生成する。ただし、 E_{max} は常数であり、 $\lambda \in (0, 1]$ となるように設定されている。

式 (4) において、 i, j は画面座標を表す。 $\mathbf{x}$ は画面座標 (i, j) が対応している 3 次元空間位置 $P(x, y, z)$ を表し、 $\mathbf{x} = [x, y, z]^T$ である。 μ_s は推定した音源の 3 次元位置を示すベクトル $\mu_s = [x_s, y_s, z_s]^T$ であり、 Σ_s は 3×3 の共分散行列である。また $d = 3$ である。

4・3 視覚・聴覚・行動の統合による注意点の選択 画像の各フレームにおいて、ロボットの行動状態、視覚、聴覚の 3 モデルから得られたマップ $Pr(i, j)$, $I_v(i, j)$, $I_a(i, j)$ を 80×60 の解像度に正規化し、 80×60 の積分発火ニューロンマップ I に入力する。各積分発火ニューロン $I(i, j)$ は上述のマップの各画素に対応しており、式 (5) に示す微分方程式に基づいて各マップの画素からの入力を積分する。ここでの q は積分を行う回数を示す。

$$\tau_1 \frac{dI(q)}{dq} = -I(q) + \gamma Pr(\alpha I_a + \beta I_v) \cdots \cdots (5)$$

$$\tau_2 \frac{dV(q)}{dq} = -V(q) + \omega I(q) \cdots \cdots (6)$$

積分値 $I(i, j)$ をさらに別の 2 次元積分発火ニューロンマップ $V(i, j)$ に入力し、式 (6) に示すような微分方程式によって積分を行い、 $V(i, j)$ の値を求める。式 (5) と式 (6) において、 α, β は視覚情報、聴覚情報の各々を重視する重みを示す。視覚、聴覚の双方を同じ割合で重視するように、 $\alpha = 0.5, \beta = 0.5$ としている。 γ, ω は積分速度を調節する重みである。積分発火ニューロンが、大きい入力に対しては発火するまで比較的少ない積分回数、小さい入力に対しては比較的多い積分回数を必要とするように、 $\gamma = 100, \omega = 0.1$ としている。 $\tau_1 = 0.3, \tau_2 = 0.3$ は積分減衰係数である。これらの値は、入力 (式 (5) の項 $Pr(\alpha I_a + \beta I_v)$) の値を 0.9 と 0.1 にした時、100 回積分を繰返す間に生成するパルス数の差が最も大きくなるように調整したものである。

あるニューロン $p(i, j)$ において、 $V_p(t) > V_{th}$ になると、このニューロンは発火する。ニューロン $p(i, j)$ が発火すると、 $\text{Dist}(p, p_k) < D$ を満たすすべてのニューロン p_k に対して、式 (7) のように V_{p_k} と I_{p_k} の値を初期値 V_0, I_0 にリセットする。 $\text{Dist}(p, p_k)$ はニューロン p と p_k との距離を求める関数である。 D はニューロン間の距離を表し、画像の解像度を考慮して、 $D = 8$ とする。また、 V_0, I_0 は初期値であり、0 とする。 V_{th} は V の閾値であり、積分速度、

減衰速度を考慮して $V_{th} = 800$ とする。 $V_{th} = 800$ の場合、各ニューロンは異なる入力に対して異なる頻度で発火するという特性を持つ。すなわち、このようなパラメータの設定により、ニューロンが大きな入力に対して先に発火することができる。

$$\begin{cases} V_{p_k}(q + \Delta q) = V_0 \\ I_{p_k}(q + \Delta q) = I_0 \end{cases} \cdots \cdots (7)$$

ニューロンマップ $V(i, j)$ 上のある場所でニューロンが発火すると、このニューロンが勝者であり、ロボットの視覚システムはこの勝者に対して反応する。ここで、勝者が現れるまでの積分回数 q は、この勝者に対応した画像領域に対するロボットの反応時間 (Response Time) を表している。ある画像フレームにおいて、一度勝者となった領域について、値 I, V をリセットした後に再び勝者となるまで、式 (5) ~ 式 (7) を繰り返して実行する。この様にすることにより、誤った領域を勝者として検出してしまうことを防いでいる。

5. 注意選択とそれに基づく行動選択の実験

式 (5) では、 $Pr = 1, I_a = 0$ となる場合、ロボットはシーンの画像情報からの刺激だけによって反応し、注意対象を選択する。しかし、ロボットが環境内で行動する際、注意選択システムが画面上の各場所からの視覚刺激 I_v に対して、すべて一様な重みで積分するのではなく、行動状態や音源位置の提示などにより場所的に偏って注意先を選択する。ここでは、ロボットの行動状態を考慮しないモデル、及び、行動状態を考慮した注意先の選択モデルを用いて、ロボットを環境内で走行させ、ロボットの移動方向上に動いているユーザや障害物があつた場合、両モデルにより選択されたロボットの注意先を比較する。ただし、実験結果を示す各図において、積分を繰り返していくにつれて発火した領域を最初から 6 番目まで表示している。

5・1 注意領域の選択 図 6 は第 1 フレームでの注意選択の結果を示す。このフレームでは、ロボットがパラメータ $\text{Speed}=2, \text{Direction}=0^\circ$ で走行している。行動状態を考慮しない場合、式 (6) への入力マップは I_v であり、図 6 の右上に示してある。マップ I_v に基づいて選択した注意先を、左上の入力画像に重畳した 6 個の で示す。特徴が顕著な画像領域、例えばエッジ強度や輝度などの高い領域に集中しており、移動方向上で障害物となる可能性がある人物に注意を向けることができていない。すなわち、ロボットは行動に関連した視覚情報を選択することができていない。

走行中のロボットにおいては、行動に緊密した画像情報、例えば障害物などを含んだ画像領域に対して、最短時間で反応できることが望ましい。すなわち、障害物領域における積分発火ニューロンは、その他の領域のニューロンと比べ、比較的短い発火時間 q で発火することが望ましい。ロボットの注意密度が実空間内で一様に分布した、行動状態

図 6 第 1 フレームでの注意先の選択結果 上段: 行動状態を考慮していない場合の選択結果, 下段: 行動状態を考慮した場合の選択結果 (a) 入力画像と注意対象の選択結果 (), (b) 行動状態, 視聴覚の 3 モデルからのマップを統合した結果

Fig. 6. Attention points selected by applying two different models separately. Top: Selection results without considering the current status of behavior. Bottom: Selection results with considering the current status of behavior. (a) Original initial frame and attention points selected. (b) Saliency maps.

を考慮しないモデルを用いる場合, 行動状態に緊密した領域での I_v が低く, 行動状態に関連しない領域での I_v が高いのであれば, ロボットの視覚システムにおいて, 前者の領域に対する反応時間 q が, 後者の領域に対する反応時間に比べて遅くなる。すなわち, この領域でのニューロンが発火するまで, より多い積分回数が必要となる。

そこで, 注意密度が場所に応じて異なるような分布を導入することによって, ロボットの行動空間での I_v を強調し, その他の領域での I_v を抑制する。各場所でのニューロンにおいて積分を行う際, ロボットの行動空間でのニューロンが先に発火する。これによって, ロボットが行動空間内の画像領域を優先的に処理することができ, 環境変動に応じて行動を素早く選択して対応することができる。

図 6 の下段には, 第 1 フレームの画像に対して, 行動状態を考慮したモデルを適用した結果を示している。右下のマップが Pr , I_v , I_a を統合したマップである。左下はシーンの画像に注意先の選択結果をで重畳して表示したものである。画面上の人物が移動中のロボットの障害物となるため, これらの注意選択結果がロボットの直進動作に直接係わっており, ロボットが行動関連の視覚情報を選択していることが分かる。

第 106 フレームにおいて, 2 つの選択モデルを用いた場合の注意選択結果を図 7 に示す。このフレームでは, ロボットが $Speed=2$, $Direction=-30^\circ$ で走行している。すなわち, ロボットは画面の左奥方向へ直進する。図 7 の左上の画像に重畳したは, 行動状態を考慮しないモデルに基づいて選択した注意先であり, 行動状態との関連性が低い。

図 7 第 106 フレームでの注意先の選択結果 ((a), (b) は図 6 と同様)

Fig. 7. Attention points selected at No.106 image frame. (Captions of (a) and (b) are the same as those in Fig.6.)

一方, 行動を考慮した注意選択モデルを用いた場合, 画面左側での注意密度が高くなるため, そこに存在する机, 人物が優先的に選択されている。

すなわち, 行動状態に応じた異なる注意密度分布を導入することにより, ロボットが行動に緊密した画像領域に対して素早く反応することができ, 環境の変化によって行動を適時に選択することが可能となる。

5.2 注意点の認識に基づく行動選択 知能ロボットにおいては, 様々に変化する環境下でユーザと共存するために, 一般に環境・ユーザの状況に応じた多くの行動モジュールが作られている。行動モジュールを選択するという行動は, ロボットの選択的注意に深く係わっている。

5.2.1 注意対象が障害物である場合 ここでは, 環境内で走行するロボットが人間や障害物に衝突しないようにしながら, 注意を向けた画像領域に対する認識に基づいて移動状態を変更させる。さらに, このように行動しているロボットに対して, 別のユーザがロボットの注意を喚起するように声をかけた場合, ロボットは元の行動を中止し, 発話したユーザとのコミュニケーションを取るように, そのユーザとの相対的な姿勢を調整する。このように注意先の動的な変更・認識という機能を付与することによって, ロボットが環境内での行動やタスクを適切に切り替えることができる。

図 2 に示すノード “Obstacle” の確率 $Pr(Obstacle = 1)$ がある閾値以上になる場合, 注意された領域は障害物領域であると認識される。図 8 は, ロボットがユーザの後方を走行するシーンを示す。この際, ロボットとユーザの移動速度の違いによって, ロボットがユーザに接近しすぎる場合がある。図 8 で, 入力画像に重畳されたは, 障害物/非障害物の認識結果を示している。障害物領域と認識された注意先を大きな円で, 非障害物領域を小さな円で示す。障害物であると認識すれば, ロボットは回避動作を選択し,

図8 注意対象に対しての認識結果 左上から順に第1, 9, 13, 17, 25, 29フレームでの障害物/非障害物の認識結果

Fig. 8. Recognition results of obstacle/non-obstacle at the 1st, 9th, 13th, 25th, and 29th frames are shown from left to right and top to bottom.

図9 連続した各フレームにおける発話中の顔の認識結果、及び、ロボット本体の位置・姿勢を調整した後のカメラから見たシーン画像(左から右, 上から下の順)

Fig. 9. Recognition results of faces that speak to the robot. Frames are shown in temporal order from left to right and top to bottom.

移動速度の減少または方向変更によって、障害物と衝突しないように動作する。

5・2・2 注意対象がユーザの顔である場合 ロボットが選択した注意領域が発話中の顔である場合、ロボットはユーザとのインタラクションを行うために、ユーザからの身体的情報を取得しやすいように、本体の姿勢と位置を調整し、話者と一定の距離や相対姿勢を保つ。ロボットが注意を向けたオブジェクトが、正面顔、音声音の音源という2つの条件を同時に満たせば、それはロボットに向かって発話する顔と認識される。

図9のシーンでは、ロボットに対して、ユーザが音信号を発している。ロボットは聴覚刺激に応じて、音源に視覚的注意を向ける。音源を視覚的注意先として選択した後、ロボットは注意を向けた領域においてオブジェクト認識を行う。図9の場合、ロボットは、注意対象が自分に向いて発話するユーザの顔であると認識し、そのユーザが自分とイ

ンタラクションしようとしていると判断する。従って、ロボットは自分の位置と姿勢を調整し、ユーザとの相対的位置関係、相対的姿勢を保つ。

6. まとめ

生物行動をまねた人間共存型ロボットの構築のためには、視覚による注意システムが重要である⁽³⁾。マルチモーダル間のクロス効果は、視聴覚センサを有した知能ロボットにおける注意システムの重要な特徴である。そこで、本論文では、まず、ベイジアンネットワークを用いて、ロボットの行動状態により空間内の各領域に分布した注意密度の大きさをモデリングした。このベイジアンネットワークにより、同一条件下での人間の注意選択結果を統計的に分析することができる。次に、注意密度の分布による注意選択への影響を考慮し、視聴覚情報を統合した注意選択モデルについて検討した。この結果、ロボットの行動に緊密した環境内の視覚的情報を優先して選択させることができた。さらに、視覚的注意の選択結果に基づいて、ロボットの行動モジュールを選択して実行させることを行った。ロボットの注意モデルの構築と実験によって、ロボットの行動状態を考慮したマルチモーダル刺激源への選択的注意モデルが、環境探索やユーザとのインタラクションを行うために重要な要素であることを明らかにした。

今後の課題としては、研究室やオフィスのような少人数のユーザグループにおいて、環境内オブジェクトやユーザに関する知識ベース、及び、ユーザとのインタラクションや環境内での行動を行う際のルールを作成して、ロボットに実装することが挙げられる。また、具体的なタスクを実現するために、ロボットの注意選択、注意先の認識、行動選択という3つの要素を一体化してロボットを自律的に行動させることを試みていきたい。

(平成16年4月30日受付, 平成16年12月27日再受付)

文 献

- (1) R. A. Brooks: "Humanoid robots", Communications of the ACM, Vol.45, Issue 3, pp.33-38 (2002-3)
- (2) B. Scassellati: "Foundations for a theory of mind for a humanoid robot", MIT Department of Computer Science and Electrical Engineering, Ph.D. Thesis (2001)
- (3) T. Fong, I. Nourbakhsh, and K. Dautenhahn: "A survey of socially interactive robots", *Robotics and Autonomous Systems*, Vol.42, No.2003, pp.143-166 (2003)
- (4) D. H. Ballard: "Animate vision", *Artificial Intelligence*, Vol.48, Issue 1, pp.57-86 (1991-2)
- (5) R. C. Arkin, M. Fujita, T. Takagi, and R. Hasegawa: "An ethological and emotional basis for human-robot interaction", *Robotics and Autonomous Systems*, Vol.42, Issue.3-4, pp.191-201 (2003-3)
- (6) B. R. Duffy: "Anthropomorphism and the social robot", *Robotics and Autonomous Systems*, Vol.42, Issue.3-4, pp.177-190 (2003-3)
- (7) M. Imai, T. Ono, and H. Ishiguro: "Physical relation and expression: Joint attention for human-robot interaction", *IEEE Trans. Industrial Electronics*, Vol.50, No.4, pp.636-643 (2003-8)
- (8) L. M. Garcia, A. A. F. Oliveira, R. A. Grupen, D. S. Wheeler,

- and A. H. Fagg: "Tracing patterns and attention: humanoid robot cognition", *IEEE Trans. Intelligent Systems*, Vol.15, No.4, pp.70-77 (2000-8)
- (9) H. D. Tagare, K. Toyama, and J. G. Wang: "A maximum-likelihood strategy for directing attention during visual search", *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.23, No.5, pp.490-500 (2000-5)
- (10) L. Itti, C. Koch, and E. Niebur: "A model of saliency-based visual attention for rapid scene analysis", *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.20, No.11, pp.1254-1259 (1998)
- (11) S. Vijayakumar, J. Conradt, T. Shibata, and S. Schaal: "Overt visual attention for a humanoid robot", *Proceeding of International Conference on Intelligence in Robotics and Autonomous Systems(IROS 2001)* (2001)
- (12) C. Balkenius and N. Hulth: "Attention as selection-for-action: A scheme for active perception", *Proceeding of EU-ROBOT'99* (1999)
- (13) T. Kimura, T. Miura, S. Doi, and Y. Yamamoto: "Top-down and bottom-up controls of attention in three-dimensional space when observers are moving forward", *Journal of Vision*, Vol.3, No.9, page 739a, [http://journalofvision.org/3/9/739/\(2003\)](http://journalofvision.org/3/9/739/(2003))
- (14) B. Chen, M. Meguro, and M. Kaneko: "Probabilistic integration of audiovisual information to localize sound source in human-robot interaction", *Proceeding of 12th IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN'2003)*, 4A4 (2003)
- (15) C. A. Seguin: "Models of neurons in the superior colliculus and unsupervised learning of parameters from time series", Ph.D. Thesis, University of Illinois at Urbana-Champaign (1998)
- (16) H. Colonius and A. Diederich: "A maximum-likelihood approach to modelling multisensory enhancement", in *Advances in neural information processing systems*, T. G. Dietterich, S. Becker, and Z. Ghahramani, Eds., Vol.14. Cambridge: MIT Press (2002)
- (17) T. Anastasio, P. Patton, and K. Belkacem-Boussaid: "Using Bayes' rule to model multi-sensory enhancement in the superior colliculus", *Neural Computation*, Vol.12, pp. 1165-1187 (2000)
- (18) J. Driver and C. Spence: "Attention and the crossmodal construction of space", *Trends in Cognitive Sciences*, Vol.3, No.7, pp.254-262 (1998)
- (19) C. B. Ferrell: "Orientation behavior using registered topographic maps", *Proceedings of the 4th International Conference on Simulation of Adaptive Behavior*, Cape Cod, pp.94-103 (1996)
- (20) L. F. Abbott: "Pointing the way toward target selection", *Nature Neuroscience*, Vol.2, No.8, pp.688-689 (1999-8)
- (21) R. Hahnloser, R. J. Douglas, M. Mahowald, and K. Hepp: "Feedback interactions between neuronal pointers and maps for attention processing", *Nature Neuroscience*, Vol.2, No.8, pp.746-752 (1999-8)
- (22) R. Mudra, R. Hahnloser, and R. J. Douglas: "Integrating neuromorphic action-oriented perceptual inputs to generate a navigation behaviour for a robot", *International Journal of Neural Systems*, Vol.9, No.5, pp.411-416 (1999)

陳

彬 (正員) 1995 年中国合肥工大卒。2001 年電気通信大学大学院電気通信学研究科機械制御工学専攻博士前期課程了。2004 年同大学院電子工学専攻博士後期課程了。同年 (株) 富士通研究所に入社、現在に至る。在学中、画像処理、知的ヒューマンインタフェース、ロボットビジョンなどの研究に従事。博士 (工学)。電子情報通信学会会員。

金

金子正秀 (非会員) 1976 年東大・工・電子卒。1981 年同大学院博士課程了。同年国際電信電話 (株) 入社、研究所に勤務。1994 年東大・工助教授。1997 年国際電信電話 (株) に戻り研究所に勤務。1998 年電気通信大学大学院電気通信学研究科助教授、2004 年同教授、現在に至る。画像符号化、3 次元画像処理、知的ヒューマンインタフェース、顔画像処理などの研究に従事。工博。電子情報通信学会、映像情報メディア学会、情報処理学会、日本 ME 学会、日本顔学会、IEEE 各会員。