自律的インタラクションを通じたマルチモーダル共同注意獲得の実現

Construction of multimodal joint attention acquisition through autonomous interaction

○ 藤木新也 (阪大院)，吉川雄一郎 (JST ERATO)，中野吏，浅田稔 (阪大院,JST ERATO)

Shinya Fujiki(Osaka Univ)，Yuichiro Yoshikawa(JST ERATO)，Tsukasa Nakano，Minoru Asada(Osaka Univ, JST ERATO)

Abstract Some robotics researchers have attended to construction of robot which acquires joint attention through interaction with human as a constructive research of human commnunication. In the previous research, robot's and human's actions were often assumed to be synchronized each other. In this paper, we propose a learning mechanism which acquires joint attention through more natural interaction in which human and robot act independently. We demonstrate that robot can acquire joint attention through a computer simulation and human-robot interaction.

1 はじめに

他者と同じものを見る行動である共同注意は, 発達心理学において, 人が他者とコミュニケーションを行うための基礎になると言われており, その発達を可能にするメカニズムの解明に注目が集まっている．共同注意は, ロボットが人とコミュニケーションする上でも, 必要な能力であると考えられ, ロボット工学においても注目されている．また, 共同注意の能力を発達させるロボットのメカニズムを考えることは, 人の共同注意発達過程の解明につながるのではないかと期待されている．従って, 共同注意の能力を発達させるロボットを実現することは上記の意味で重要なことであるといえる．

これまで, 人の共同注意の発達を知るために, ロボットを用いた構成論的アプローチ [1] による研究が行われてきた．長井ら [2] はロボットが共同注意を行うために必要となる視線追従機能を獲得する学習モデルを提案し, インタラクションを通じ, 人に明示的な教示をされることなく視線追従が獲得できることを示した．一方, 人の乳児は共同注意に視線だけでなく，発話も利用していると考えられる．中野ら [3] は, 明示的な教示がない状況での視線追従機能と言葉（物体の名称）の同時学習モデルを提案し, 視線追従機能の学習と物体の名称の学習が相互促進可能であることを示した．しかし, これらの研究においては, 人とロボットの同期性, すなわち人が必ずロボットに対して先行して視線を動かすことがが仮定されていた．

本研究では, より現実的な状況下での発達の課題を取り扱うため, 人やロボットが相手に対し同期的にふるまうとは限らない状況におけるインタラクション (自律的インタラクション) を想定する．同様な状況での共同注意の獲得を扱った研究として, Sumioka et al.[4] や Triesch et al.[5] の研究がある．Sumioka et al. が指摘したように自律的インタラクションを通じた学習では, ロボットと人の視線移動のタイミングの違いにより, 人の視線と自身の視線移動の誤った対応関係が学習データとして入力されるという問題がある．これらの従来研究では, 視線追従機能のみの学習が取り扱われていたが, 中野ら [3] のモデルのように視線追従機能と言葉の同時学習を行うことで, 共同注意に視線と言葉という複数の情報を利用できるようになるため, 人の行動とのタイミングを合わせやすくなり誤った対応関係の学習が抑えられることが期待できる. 本研究では, 中野ら [3] のシステムを拡張し, ロボットが人との自律的インタラクションを通じてマルチモーダルな共同注意獲得の実現を目指す．

本稿では, 2 節で本研究における共同注意の問題設定を説明し, 3 節で提案システムについて説明する．そして, 4 節で計算機シミュレーションとロボットを使った人とのインタラクション実験について述べる. 5 節でまとめと今後の課題について述べる．

2 問題設定

図 1 に示すように, 人とロボットがテーブルを挟んで向かい合う状況を想定する．テーブルは N 個の領域に区画されているとし, そのうちの M_0 個の物体が各区画に一つずつ配置されているとする．人はロボットあるいは物体を注視し, 物体注視の前後に時折，その物体の名称を発話する．ロボットのタスクは, そのような人の発話や注視行動を知覚し, 自らも物体を注視をするという経験を通じ, 人の顔とそれが指す場所の対応関係と言葉とそれが指す物体の対応関係を学習すること, すなわち, 人が注目しているのと同じ対象物を注視できるようにな

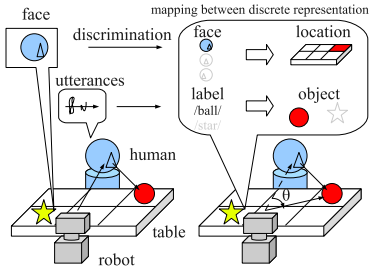


図 1: 共同注意の学習を行う環境

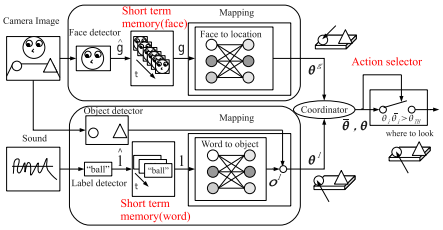


図 2: 提案システム

ることである．ここで, 従来研究 [2, 3] と異なり, 人の行動表出のタイミングは, ロボットのそれに対して同期的であるとは限らない, すなわち, 人が必ずロボットに先行して視線を動かすことは仮定していない．

3　共同注意メカニズム

図 2 に提案システムの概略を示す．自律的インタラクションを通じた学習に対応させるため, 中野ら [3] のシステムに, 人の視線や発話に関する情報を保持する短期記憶 (Short term memory) と視線を切り替えるタイミングを決定する行動選択器 (Action selector) を追加した．

3.1　注視メカニズム

3.1.1　視線による注意モジュール

視線による注意モジュール (図 2 上部) には, はじめにカメラ画像が入力される．顔検出器 (Face detector) により, 1 フレーム毎に人の顔の方向についての観測情報を表すベクトル $\hat{\boldsymbol{g}}$ が求められる．ここで, 人の顔は常にロボットの視野内にあると仮定し, 顔検出器により 30Hz のサンプリング周期で人の顔の検出が可能であるとする．人の注視方向の違いは，検出された顔画像の違いとして表れると仮定し，システムは人がN個のテーブルのそれぞれの区画を注視しているときの顔画像と，正面（ロボット）を見ているときの顔画像を識別できるとする．$\hat{\boldsymbol{g}}$ は検出された顔が予め登録された N＋1 通りの顔画像のいずれに近いものであるかの識別を表す N＋1 次元ベクトルであり，観測された人の顔に対応する要素に 1, その他の要素には 0 が格納される．$\hat{\boldsymbol{g}}$ は短期記憶に送られしばらくの間保持され, 各フレームで短期記憶から過去 R_{g1} フレーム分の $\hat{\boldsymbol{g}}$ の和である $N+1$ 次元ベクトル $\tilde{\boldsymbol{g}}$ が計算される．この要素 $\tilde{g}_i$ を $g_i = \tilde{g}_i / \sum_k \tilde{g}_k$ により大きさが 1 になるように正規化して得られたベクトル $\boldsymbol{g}$ が短期記憶から出力される．

顔情報に関するベクトル $\boldsymbol{g}$ は顔-場所マッピングに入力され, 共同注意のために注視されるべき位置が $\boldsymbol{\theta}^g = W^g \boldsymbol{g}$ のように出力される．ここで, $\boldsymbol{\theta}^g$ は人の視線を手がかりとしたとき，テーブルの各区画が注視されるべき程度を要素とする N 次元ベクトルである．従って, 顔-場所マッピングの結合荷重である W^g は $N \times (N+1)$ 行列である．正面 (人) への注視行動については, 3.1.3 でも説明するように特定の状況でマッピングの出力に関係なく行われる．$\hat{\boldsymbol{g}}$ の認識では, ロボットは人の顔の向きの違いが分かるのみであり, 人が見ている所を見れるようになるためには, 適切な W^g を学習する必要がある．

3.1.2　言葉による注意モジュール

言葉による注意モジュール (図 2 下部) には，はじめに音声が入力される．ラベル検出器 (Label detector) により, 1 フレーム毎に人が発した物体のラベルについての観測情報を表す $\hat{\boldsymbol{l}}$ が求められる．$\hat{\boldsymbol{l}}$ は人が L 種類のラベルのいずれを発したかを表す L 次元ベクトルであり, 観測されたラベルに対応する要素に 1, その他の要素には 0 が格納される．$\hat{\boldsymbol{l}}$ は短期記憶に送られしばらくの間保持され, 各フレームで短期記憶から過去 R_{l1} フレーム分の $\hat{\boldsymbol{l}}$ の和である L 次元ベクトル $\tilde{\boldsymbol{l}}$ が計算される．$\tilde{\boldsymbol{l}}$ を $\tilde{\boldsymbol{g}}$ と同様に正規化して得られたベクトル $\boldsymbol{l}$ が短期記憶から出力される．

ラベル情報に関するベクトル $\boldsymbol{l}$ は言葉-物体マッピングに入力され, 共同注意のために注視されるべき物体を $\boldsymbol{o}^l = W^l \boldsymbol{l}$ のように出力する．ここで，$\boldsymbol{o}^l$ は人の発話を手がかりとしたときに各物体が注視されるべき程度を要素とする M 次元ベクトルである．また W^l は言葉-物体マッピングの結合荷重であり, $M \times L$ 行列である．人の発話を手がかりとしたときにテーブルの各区画が注視されるべき程度を要素とする N 次元ベクトル $\boldsymbol{\theta}^l$ の要素 ${\theta_i}^l$ には, 番号 i の区画に番号 j の物体が存在すれば ${o_j}^l$ が代入され, そうでなければ 0 が代入される．ここで, ロボットは環境中のすべての物体が見えていると仮定する．$\hat{\boldsymbol{l}}$ の認識では, 人の発話の違いが分かるのみであり, 発話されたラベルが指すものを見れるようになるためには, 適切な W^l を学習する必要がある．

3.1.3 モジュールの統合と自信に基づく注視メカニズム

2 つの注意モジュールの出力 $\boldsymbol{\theta}^g$ と $\boldsymbol{\theta}^l$ を統合して，注視対象が決定される．決定には統合されたベクトル

$$\boldsymbol{\theta} = \frac{\sum_{k\in\{g,l\}} \theta^k_{max}\boldsymbol{\theta}^k}{\sum_{m\in\{g,l\}} \theta^m_{max}} \tag{1}$$

の要素 θ_i を $\bar{\theta}_i = \theta_i / \sum_k \theta_k$ により正規化して得られたベクトル $\bar{\boldsymbol{\theta}}$ を用いる．ここで,θ^k_{max} は $\boldsymbol{\theta}^k$ の要素の最大値である．$\boldsymbol{\theta}^k$ は大きさが 1 となるよう正規化されたベクトルであるため，θ_{max} が高い値であることは，特定の要素のみが選択された状態であることを表す．従って, 式 (1) の統合により, 確信度の高いモジュールの出力に重みをつけて利用できる．$\bar{\boldsymbol{\theta}}$ の計算は視線移動時を除いて毎フレーム行われ, この各要素の値に対応するテーブルの区画を注視する確率として, 注視すべき区画の選択に利用される．

また, $\boldsymbol{\theta}$ および $\bar{\boldsymbol{\theta}}$ は視線移動を行うか否かの判定にも使われる．選択された区画の番号を j^* とすると, $\bar{\theta}_{j^*}\theta_{j^*} > \theta_{TH}$ が満たされれば注視動作を開始する．また, 現在の区画を注視し初めてから R_0 フレームが経過すると注視動作が開始される．このとき, その視線の移動先は, 現在見ている場所が正面（人）の時はテーブルの区画 j^* とし, テーブルの時はマッピングの出力に関係なく正面とした．ここで, 正規化された $\bar{\theta}_{j^*}$ は区画 j^* について, その他の区画を見ることに対する相対的な自信の大きさ表し,θ_{j^*} は区画 j^* 自体を見ることの自信の大きさとみなすことができる．よって, これらの積 $\bar{\theta}_{j^*}\theta_{j^*}$ が閾値 θ_{TH} を超えたときにその区画を見ることへの自信があるとし, 注視動作を開始する．

ロボットが学習の進度に応じた注視動作開始の決定を行えるようにするため,$\bar{\theta}_{j^*}\theta_{j^*}$ の値の履歴に応じて θ_{TH} を変化させる．具体的には, $\theta_{TH} = m_\theta + a\sigma_\theta$ により計算される．ここで m_θ, σ_θ は, 過去にテーブルの j^* 番目の区画を注視すべきと判定されたときの出力 $\bar{\theta}_{j^*}\theta_{j^*}$ の平均と標準偏差であり, a は定数である．この注視メカニズムにより, 学習が進むと人の注視行動に応答しやすくなり, ロボットがより多くの注視動作をすることが期待できる．

3.2 学習メカニズム

前項の注視メカニズムにしたがって注視したテーブルの区画, 物体のパターンおよび, 短期記憶に保持されている人の観測情報をもとに図 2 の 2 つのマッピングを学習させる．ただし, 人とロボットは各々で自律的インタラクションを行うため, ロボットが取得する学習データの中に誤った対応関係が多く含まれるという問題が生じる．この問題に対処するため，中野ら [3] の用いた相互排他性に基づく学習原理を適用する．相互排他性は，発達心理学において幼児が語彙を獲得する際のバイアスとして知られており [6], 中野ら [3] のシステムでは複数の 1 対 1 マッピング学習での相互促進可能な学習原理として利用されている．

周囲の物体やその位置が変化する様々な状況下において，学習すべき対応関係は相互排他的であるのに対し，それ以外の対応関係は偶然見つかるものの，異なる状況にわたって一貫するものでない．そのため，相互排他性を考慮した学習則を用いることでロボットは正しい対応関係を学習できると期待できる．学習則には, 中野ら [3] の用いた交差投錨型ヘッブ則 [7] を修正したものを適用した．

注視行動の後，物体を発見した場合，ロボットは以下のようにマッピングの結合荷重を更新する．人の視線と注視したテーブルの区画の対応づけの学習 (図 2 上部のマッピングの学習) では，まず短期記憶によって，ロボットの注視動作開始の R_{1g} フレーム前から，注視動作終了の R_{2g} フレーム後までの間の顔情報が $N+1$ 次元ベクトル $\boldsymbol{g}$ として出力される．そして，$\boldsymbol{g}$ に従って確率的に選択された入力要素 i^* と注視した区画の ID に対応する出力要素 j^* の結合荷重 $w^g_{i^*j^*}$ が次式によって更新される．

$$w^g_{i^*j^*}(t+1) = w^g_{i^*j^*}(t) + \xi_g\mu_g(1.0 - w^g_{i^*j^*})(t) \tag{2}$$

ここで ξ_g は学習率を調整する定数であり，$\mu_g(=\eta^g_{ij}\times\eta^\theta_{ij})$ は要素 i^* と要素 j^* 間の相互排他性を表現したものである．また η^g_{ij}, η^θ_{ij} はそれぞれマッピングの入力要素と出力要素の排他度を表しており，以下の式で計算される．

$$\eta^g_{ij} = \exp\left(-\frac{\sum_{k,k\neq j} w^g_{ik}(t)}{\alpha^2_g}\right) \tag{3}$$

$$\eta^\theta_{ij} = \exp\left(-\frac{\sum_{k,k\neq i} w^g_{kj}(t)}{\alpha^2_\theta}\right) \tag{4}$$

同時に, $w^g_{i^*j^*}(t)$ 以外の結合荷重は側抑制によって

$$w^g_{ij^*}(t+1) = w^g_{ij^*}(t) - \beta_g(1-\eta^g_{ij^*})\Delta w^g_{i^*j^*}(t) \tag{5}$$

$$w^g_{i^*j}(t+1) = w^g_{i^*j}(t) - \beta_\theta(1-\eta^\theta_{i^*j})\Delta w^g_{i^*j^*}(t) \tag{6}$$

のように減らされる．β_g は減衰率を調整する定数であり，$\Delta w^g_{i^*j^*}(t)$ は式 (2) の右辺第 2 項である．

また物体を発見できなかった場合には, 入力要素 i^* と出力要素 j^* の結合荷重 $w^g_{i^*j^*}$ は次式で更新される．

$$w^g_{i^*j^*}(t+1) = b_g w^g_{i^*j^*}(t) \tag{7}$$

ここで, b_g は減衰率を調整する [0,1] の定数とし，誤った対応関係として結合荷重を抑制する．

人の発話ラベルと物体の対応づけの学習 (Fig.2 下部のマッピングの学習) において発話ラベル $\boldsymbol{l}$ からの i^* の求め方, マッピングの更新則は視線の学習の場合と同様である．顔-場所マッピングの学習パラメータ R_{1g}, R_{2g}, b_g に相当するパラメータは R_{1l}, R_{2l}, b_l である．また物体が発見できたときには, その物体の ID を j^* とし, 発見できなかったときには j^* は言葉-物体マッピングの出力 $\boldsymbol{o}^l$ を正規化したものから確率的に求める．

提案システムでは中野ら [3] と同様に確信度の高い方のモジュールに従って動くことで，2 つのモジュールの効率的な学習ができる．さらに，本システムでは確信度に応じて注視動作開始を決定する注視メカニズムを採用することで，誤認識や人の注視移動中のセンサデータをはじくことができ，効率的な学習ができると期待される．

4　インタラクション実験

提案システムによりロボットが自律的なインタラクションでの共同注意獲得が可能であるかを検証した．はじめに提案システムを実装した実ロボットと人とのインタラクション実験を行い, 学習可能であるかを調べる．その後, 計算機シミュレーションにより視線と語彙の同時学習が共同注意の学習にどのような影響を与えるかを検証する．

4.1　実ロボットを用いた実験

4.1.1　実験設定

本実験では, 子供型ヒューマノイドロボット CB^2(Child robot with Biomimetic Body)[8] を用いて実験を行う．実験環境を図 3 に示す．ロボットと被験者はテーブルを挟んで向かい合っている．テーブルは 3 つの区画 (左, 真ん中, 右)($N = 3$) に区切られている．被験者には 3 つの物体 (くるま, ボール, バット)($M = 3$) が渡されており, そのうち 2 つが被験者によってテーブルの各区画に重複なく配置される ($M_o = 2$)．被験者はロボットに物体の名前を教えるため, テーブルの上の物体やロボットの顔を注視し, 時折発話する．被験者の発話は,「くるま」,「ボール」,「バット」,「これ」の 4 種類 ($L = 4$) とし, 発話タイミングは特に制限していない．テーブル上の物体の種類および配置はおよそ 1 分ごとに変化させるよう指示した．約 20 分間のインタラクションを 3 人の被験者について行った．結果は 3 人の被験者に対する結果を平均したもので評価した．

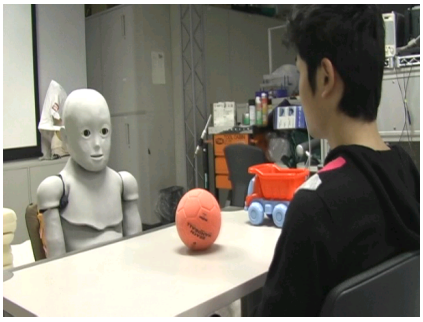


図 3: 実験環境

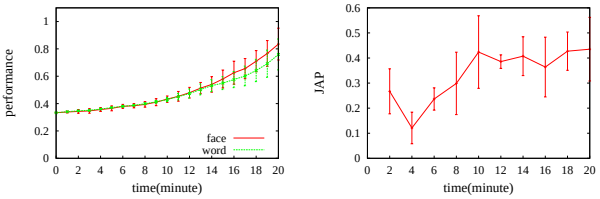
視線についての入力ベクトル $\hat{\boldsymbol{g}}$ は, あらかじめ定めた顔の角度によって人の顔を 4 種類の顔 (正面顔, 左顔, 真下顔, 右顔) に識別可能な画像処理により求めた．物体の認識は 2 次元カラーヒストグラムによるテンプレートマッチングにより行った．言葉についてのベクトル $\hat{\boldsymbol{l}}$ は大語彙連続音声認識エンジン Julius を用いて求めた．

実験に用いたパラメータを以下に示す．時間に関するパラメータは, $R_{g1} = 30, R_{l1} = 60, R_{g2} = R_{l2} = 60$ とし, R_0 は [40,80] のランダムな数値でロボットが行動を選択する度に変化させた．マッピングの結合荷重の更新に関しては, $\alpha_g = \alpha_\theta = \alpha_l = \alpha_o = 1.0$, $\beta_g = \beta_\theta = \beta_l = \beta_o = 1.0$, $\xi_g = \xi_l = 0.02$, $b_g = 0.95$, $b_l = 0.98$ とした．

4.1.2　実験結果

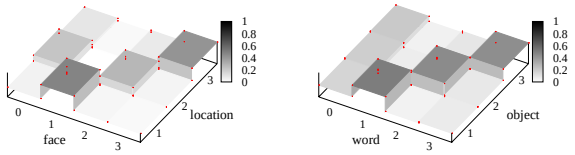
図 4(a) に視線と言葉のマッピングの習熟度の遷移を示す．視線のマッピングの習熟度は，マッピングの結合荷重を用いて $\sum_{k=1}^{N} \tilde{w}^g_{kk}/N$ で評価した．ここで, $\tilde{w}^g_{ij}$ は重みを $w^g_{ij}/\sum_{k=1}^{N} w^g_{ik}$ により正規化したものである．言葉の方も N を M として同様に評価した．図 4(a) から視線のマッピング (face) および言葉のマッピング (word) の両方が正しい対応関係を学習できていることがわかる．また，図 5(a)，(b) にそれぞれ 20 分間のインタラクション後の視線と言葉のマッピングの結合荷重の値を示す．顔 (face) の ID は 1,2,3 はそれぞれ左，真ん中，右に対応し，テーブルの区画 (location) の ID はその区画を注視する顔の ID と一致させた．言葉 (word) の ID も同様に，1,2,3 はそれぞれ「くるま」,「ボール」,「バット」に対応し，物体 (object) の ID は対応する言葉の ID と一致させた．なお，正面顔および「これ」の ID は 0 である．図 5(a)，(b) からが ID が一致する入出力間の結合荷重の値が大きく，また対応関係をもたない入力要素 (face0 および word0) は特定の出力要素と大きな結合をもたないことから，提案システムにより正しい対応関係が学習できていることがわかる．

さらに，共同注意成功率として次式で表わされる JAP



(a) マッピング習熟度の変化　　(b) JAP の変化

図 4: 学習パフォーマンス



(a) 顔-場所マッピングの結合荷重　　(b) 言葉-物体マッピングの結合荷重

図 5: 学習後の結合荷重

(joint attention performance) の遷移を図 4(b) に示す.

$$JAP = \frac{\text{過去 2 分間に人と同じものを見た回数}}{\text{過去 2 分間にテーブルへの視線移動をした回数}} \quad (8)$$

図 4(b) より JAP が徐々に増加し，10 分で 4 割程度共同注意できていることが読み取れる．ただし，JAP の最大値が 0.4 程度にしか達しない理由として，主に左右に首を振ったときに観測される人の顔画像が歪んでいるために顔認識の成功率が低下してしまったことが考えられる．また，JAP が収束している 10 分での図 4(a) の習熟度の値は 0.4 を越える程度と一見低いように思われるが，提案した注視メカニズムによって相互排他的な対応関係を発見でき，適切な注視動作開始の決定ができていると考えられる．

4.2 計算機シミュレーションによる同時学習の効果の検証

ここでは人の視線と発話という複数の情報を利用することで, ロボットが視線や言葉の対応関係の学習にどのような影響を与えているかを検証した．試行回数を増やし，定量的データでの学習パフォーマンスの比較を行うため, 計算機シミュレーションにより検証を行った．

4.2.1 環境設定

人とロボットのインタラクションを行う環境を図 6 に示す．人とロボットはテーブルをはさんで向かい合っており，テーブルは 10 分割 ($N = 10$) され，各区画にはそれぞれ ID が割り振られている．人の視線は点 O を原点とするベクトル $\boldsymbol{r}_g = (x_g, y_g)$ で表され, ベクトルの

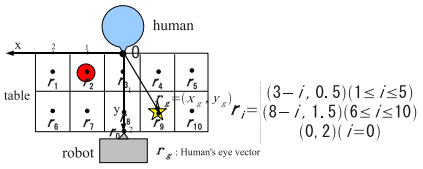


図 6: シミュレーション環境

先端を人が注視している区画とする．各区画の大きさは 1.0×1.0 とし, 人が i 番目の区画を見ているときの視線ベクトル $\boldsymbol{r}_g = \boldsymbol{r}_i$ は, 図 6 に示す通りである．$\boldsymbol{r}_0$ はロボットを見ているときの視線ベクトルである．また 10 個 ($M = 10$) ある物体が用意され，環境中にはそのうち 5 個 ($M_o = 5$) の物体が配置される．

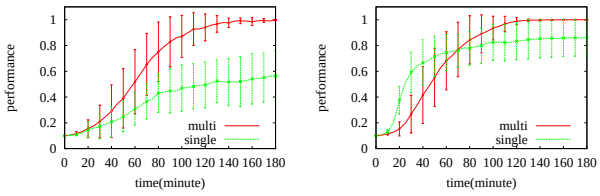
4.2.2 インタラクション設定

人はロボットと物体を交互に注視する．注視時間および注視位置への移動にかかる時間はそれぞれ C_0 と C_1 フレームとし，C_0 と C_1 はそれぞれ [40,80]，[10,20] の間で毎回ランダムに変化させた．区画 i に存在する物体を注視する場合，ロボットから区画 i への人の視線切り替えは $\boldsymbol{r}_g = \boldsymbol{r_0} + (\boldsymbol{r}_0 - \boldsymbol{r}_i)t/C_1$ に従って視線ベクトルを変化させた．また, 人は物体 (物体がある区画の中心) の注視開始の 30 フレーム前から注視終了の 30 フレーム後の間に発話する．そして，0.7 の確率でその物体の名称を発話し，0.3 の確率で関係のない言葉を発話するとした．関係のない言葉は 5 個用意し, 人の発話する言葉は全部で 15 個 ($L = 15$) である．ただし, 発話に要する時間は考慮していない．人はこの行動を繰り返し, 10 回繰り返す毎に環境中の物体の種類と配置をランダムに変化させた．

ロボットが取得する人の顔情報は, 注視 $\boldsymbol{r}_g$ に対応する区画の ID とし, $\boldsymbol{r}_g = \boldsymbol{r}_0$ のときは 0(正面顔) とした．ロボットの注視時間, 学習パラメータは実機実験と同様の値を用い，ランダムに視線を切り替える時間は人の C_1 と同様に決定させた．

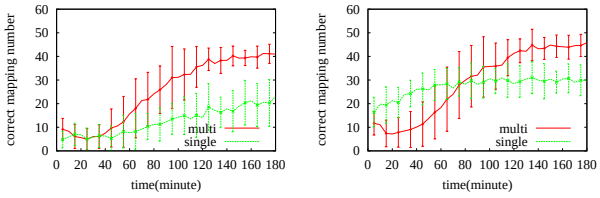
4.2.3 シミュレーション結果

計算機シミュレーションでは, 180 分程度に相当するインタラクションを 20 回行い, その平均で共同注意の結果を評価する．顔-場所および言葉-物体マッピングの習熟度の遷移を図 7 に示す．図 7 より，視線や言葉の単一の学習を行った場合 (single) に比べ，視線と言葉の同時学習を行った場合 (multi) の方が速く学習できていることが読み取れる．このことから, 自律的インタラクションにおいても, 中野ら [3] の研究で示されている視線と



(a) 顔-場所マッピング　　(b) 言葉-物体マッピング

図 7: 各マッピングの習熟度の遷移



(a) 顔-場所マッピング　　(b) 言葉-物体マッピング

図 8: 正しい対応関係の学習の回数

言葉の学習の相互促進作用により学習が加速することが確認できた.

加えて, 自律的インタラクションにおいて, 正しい対応関係学習の機会が相互促進作用によりどの程度増加しているかについて検証を行った. 図 8 に過去 5 分における視線, 言葉それぞれについて正しい対応関係を学習できた回数, すなわち, 式 (2) によるマッピングの更新時に正しく対応する i^* と j^* が選択された回数の遷移を示す.

図 8 より, 同時学習を行った場合 (multi) の方が, 視線, 言葉ともに正しい対応関係の学習を行った回数の増加が大きく, 学習が進むと視線や言葉の単一学習を行った場合 (single) の回数よりも大きくなることが読み取れる. この理由として, 1 つ目に同時学習の場合はより確信度の高いモジュールの出力を注視場所の決定に利用することで, 人と同じものを見れるようになりやすいことがあげられる. 2 つ目に視線と言葉という複数の情報がある中で, いずれかでも確信度が高ければ注視動作を開始するしくみを採用した結果, 正しい対応関係の学習を観測する経験が増えたことが原因ではないのかと考えられる. 1 つ目は中野ら [3] の研究でも指摘されていたが, 2 つ目はインタラクションを自律的にしたことにより生じた効果であるといえる. このように正しい対応関係の学習を行う割合が増えたことと, 学習の機会自体が増えたことにより, 図 7 に見られるような視線と言葉の同時学習による相互促進効果が生じたと考えられる.

5　おわりに

本稿では, 中野ら [3] のシステムを拡張し実ロボットに実装することで, ロボットが人との自律的インタラクションを通じて, マルチモーダル共同注意を獲得できることを示した. また, 計算機シミュレーションにより, 視線と言葉の同時学習は, 自律的インタラクションにおいては正しい対応関係の学習の機会を増やし, 学習パフォーマンスの向上につながることを示した.

提案システムは学習が進むにつれて人の視線にすばやく反応し, 人と同じものを見ることで正しい言葉の対応関係を学習できるようになる. そのため, 提案システムが人の乳児が言葉を覚えるときに養育者の視線情報を利用するようになる [9] という発達過程の一側面を再現している可能性があるといえる. しかし, 人の乳児は単に養育者の視線を追従するようになるだけではない. 乳児は養育者に見てほしい対象が存在するときには, 養育者の視線を誘導するようになる. 自律的に振る舞う養育者の視線を追従し, また自分の視線を追従させるメカニズムを考えることは, 人の共同注意の発達のしくみを理解する上で, 加えて, より効率的に養育者から言葉を学習するコミュニケーションロボットを実現する上で重要な課題だと考えられる.

参考文献

[1] Minoru Asada, Koh Hosoda, Yasuo Kuniyoshi, Hiroshi Ishiguro, Toshio Inui, Yuichiro Yoshikawa, Masaki Ogino, and Chisato Yoshida. Cognitive developmental robotics: a survey. *IEEE Transactions on Autonomous Mental Development*, Vol. 1, No. 1, pp. 12–34, 2009.

[2] 長井志江, 細田耕, 森田章生, 浅田稔. 視覚注視と自己評価型学習の機能に基づくブートストラップ学習を通した共同注意の創発. 人工知能学会論文誌 = Transactions of the Japanese Society for Artificial Intelligence : AI, Vol. 19, , 2004.

[3] 中野史, 吉川雄一郎, 浅田稔, 石黒浩. 相互排他性に基づくマルチモーダル共同注意. 日本ロボット学会誌, Vol. 27, No. 7, pp. 814–822, 2009.

[4] Hidenobu Sumioka, Koh Hosoda, Yuichiro Yoshikawa, and Minoru Asada. Acquisition of joint attention through natural interaction utilizing motion cues. *Advanced Robotics*, Vol. 21, No. 9, pp. 983–999, 2007.

[5] Triesch Jochen, Teuscher Christof, Deak Gedeon, and Eric Carlson. Gaze following : why (not) learn it. *Developmental Science*, Vol. 9, No. 2, pp. 125–147, 2006.

[6] Markman E and Wachtel G. Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, Vol. 20, pp. 121–157, 1988.

[7] Yuichiro Yoshikawa, Koh Hosoda, and Minoru Asada. Unique association between self-occlusion and double-touching towards binding vision and touch. *Neurocomput.*, Vol. 70, No. 13-15, pp. 2234–2244, 2007.

[8] Takashi Minato, Yuichiro Yoshikawa, Tomoyuki Noda, Syuhei Ikemoto, Hiroshi Ishiguro, and Minoru Asada. Cb2: Child robot with biomimetic body for cognitive developmental robotics. *Proceeding of IEEE-RAS International Conference on Humanoid Robots*, 2007.

[9] Baldwin D. Infants' contribution to the achievement of joint reference. *Child Development*, Vol. 62, No. 5, pp. 875–890, 1991.