相互排他性原理に基づくマルチモーダル共同注意

○中野吏 (大阪大学) 吉川雄一郎 (JST ERATO) 浅田稔 (大阪大学，JST ERATO) 石黒浩 (大阪大学，JST ERATO)

Multimodal joint attention based on mutual exclusivity principle

*Tsukasa Nakano (Osaka University), Yuichiro Yoshikawa (JST ERATO), Minoru Asada (Osaka University, JST ERATO), Hiroshi Ishiguro (Osaka University, JST ERATO)

Abstract— Developmental processes of each function through joint attention are discussed. In this paper, we study the mutual facilitation of simultaneously learning multi-functions through multimodal joint attention: gaze-driven attention and word-driven attention.We enhance the well-known mutually exclusivity bias in child language development as a general bias to simultaneously learn mappings for different functions that are expected complementary to each other, that is mutually exclusivity selection principle (μX principle). In computer simulation, we analyzed the effects of μX principle in learning multi-functions and argued the correspondence of the synthesized development to infants.

Key Words: Joint attention, Mutual exclusivity, Simultaneous learning of multi-functions, Mutual facilitation

1. はじめに

近年様々なコミュニケーションロボットが開発され，実社会において人間と共に活動することが期待されている．事前に全ての知識をロボットに与えることは困難であり，コミュニケーションロボットは人とのコミュニケーションを通じて学習を行う必要がある．

人の幼児は共同注意を通して同時並行的に語彙や視線追従能力の学習を行うことが発達心理学において知られている [1][2]．共同注意とは他者と同一の対象に注意を向ける行動であり，複数のモダリティの情報（視線，言葉，指差し等）を利用することで達成される行動である．Butterworth らは人の幼児が 12ヶ月から 18ヶ月にかけて視線追従可能な領域を拡大させているという段階的発達過程を明らかにした [1]．また Baldwin は 18ヶ月頃までに養育者の視線情報を利用して語彙を学習するようになることを明らかにした [2]．

最近では共同注意を通じて語彙あるいは視線追従のマッピングをロボットに学習させることで，人の発達過程を構成的に理解しようとする研究がある [3][4][5][6][7]．しかし，これらの研究では単一モダリティの学習に焦点をあてており，マルチモーダルな共同注意による複数のモダリティの学習は考慮されていない．一方，発達心理学で示される幼児の発達過程 [1][2] はマルチモーダルな共同注意を通して同時期に獲得される能力の相互促進的学習がもたらす結果であると考えられる．しかし，マルチモーダルな共同注意の獲得メカニズムおよびその発達過程の詳細は明らかにされていない．

そこで本研究では共同注意を通じて語彙および視線追従能力のマッピングが同時獲得可能な学習モデルを提案することで，相互促進的学習が可能なマルチモーダルシステムの構築を試みる．また本研究を通じて幼児の共同注意発達過程の理解を深める．共同注意に利用可能な情報（マッピングの学習結果）のモダリティ選択に語彙獲得における相互排他性バイアス [8]（1 対 1 マッピングのバイアス）の概念を一般化した相互排他性選択原理 (*mutual exclusivity selection principle*; 以下 μX 原理) を用いる．μX 原理はモジュール内およびモジュール間の両レベルで用いられる．モジュール内の μX 原理はマッピングの入出力の相互排他度を評価するために用いられ，モジュール間の μX 原理は各マッピングの出力の相互排他度を評価するために用いられる．

本稿では次章でマルチモーダルな共同注意システムへの μX 原理の実装方法について記述した後，2 つの実験条件における計算機シミュレーションの結果を示す．1 つ目の実験では μX 原理を適用したマルチモーダルな共同注意システムによる学習の相互促進作用を示す．2 つ目の実験では養育者が学習者の学習状況に応じて行動を変化させることで学習がさらに加速されることを示す．最後にシミュレーションと幼児の発達過程との一致および今後の問題について議論する．

2. μX 原理に基づく共同注意

ここでは視線および言葉による共同注意（以下 JA）に注目し，マルチモーダルな JA のシステムを構築する．学習者と養育者の周りには複数の対象物が存在し，学習者は養育者とのインタラクションを通して JA を獲得する．養育者は試行毎にある対象物を注視し，その名称（以下ラベル）を発話する（Fig.1 参照）．そして学習者は養育者の顔パターンと養育者の視線方向，および養育者の発話ラベルと対象物の視覚パターンとの関係を学習し，その学習結果に基づいて注視行動を決定する．学習者は JA 成功の是非に関わらず，対象物を見た場合各注意モジュールの重みを更新する．

JA 獲得のため，学習者は視覚情報 $\boldsymbol{x}_v \in \Re^{N_v}$ および音声情報 $\boldsymbol{x}_a \in \Re^{N_a}$ を観測し，注視位置 $\boldsymbol{\theta} \in \Re^{N_\theta}$ を決める．ここで N_θ, N_v および N_a はそれぞれ注視位置，視覚および音声情報の次元を表す．本システムは Fig.2 のようなマッピングへの入力を離散的に表現するセグメンテーション能力をすでに獲得していると仮定す

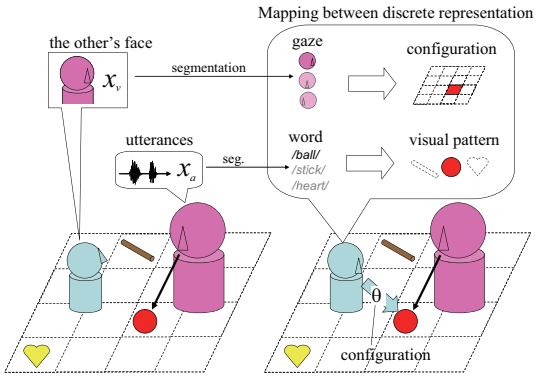


Fig.1 Situation of learning multimodal joint attention

る．ただし，このセグメンテーション能力はエラーを含む能力としてモデル化されている．$\bar{\boldsymbol{r}} \in \Re^{M_r}$ および $\bar{\boldsymbol{g}} \in \Re^{M_g}, \bar{\boldsymbol{w}} \in \Re^{M_w}, \bar{\boldsymbol{p}} \in \Re^{M_p}$ はそれぞれ注視位置および顔パターン，発話ラベル，視覚パターンのベクトルを表す．M_r と M_g, M_w, M_p は各ベクトルの次元である．観測値から $\bar{\boldsymbol{g}}$ および $\bar{\boldsymbol{w}}$ の離散ベクトルはセグメンテーション関数 ${}^{\bar{g}}\boldsymbol{S}_v : \Re^{N_v} \rightarrow \Re^{M_g}$ および ${}^{\bar{w}}\boldsymbol{S}_a : \Re^{N_a} \rightarrow \Re^{M_w}$ によって求められる．バーのついたベクトルは Fig.2 のような離散的に表現された尤度の集合を表す．

2·1 マルチモーダル注意モジュール

提案する学習機構は $\bar{\boldsymbol{g}}$ に基づく視線による注意と $\bar{\boldsymbol{w}}$ に基づく言葉による注意の二つの注意モジュールをもつ．学習者は各モジュールから得られた注視位置 ${}^{\bar{r}}\bar{\boldsymbol{r}}_g$ と ${}^{\bar{r}}\bar{\boldsymbol{r}}_w$ を $\boldsymbol{\theta}$ に統合して，その注視確率に従って注視行動を行う．各モジュールは視覚および言葉による JA を獲得するためそれぞれ注視位置ベクトル ${}^{\bar{r}}\bar{\boldsymbol{r}}_g \in \Re^{M_r}$ および ${}^{\bar{r}}\bar{\boldsymbol{r}}_w \in \Re^{M_r}$ への出力を学習する．

2·1.1 視線による注意

視線による注意モジュールはセグメンテーション関数 ${}^{\bar{g}}\boldsymbol{S}_v$ およびマッピング関数 ${}^{\bar{r}}\boldsymbol{M}_{\bar{g}} : \Re^{M_g} \rightarrow \Re^{M_r}$ によって構成される（Fig.3 上部）．${}^{\bar{r}}\boldsymbol{M}_{\bar{g}}$ は二層のネットワークで構築され，i 番目の入力ノードと j 番目の出力ノードの重みは ${}^{\bar{r}}w_{\bar{g},ij}$ で表現される．この重みは入出力の相関関係を表す．${}^{\bar{r}}\boldsymbol{M}_{\bar{g}}$ は観測された顔パターン $\bar{\boldsymbol{g}}$ から想起される注視位置 ${}^{\bar{r}}\bar{\boldsymbol{r}}_g$ を出力し，注視経験を通して重みが更新される．

入力 $\bar{\boldsymbol{g}}$ が与えられると j 番目の出力ノードの活性 ${}^{\bar{r}}a_{\bar{g},j}$ は ${}^{\bar{r}}a_{\bar{g},j} = \sum_i^{M_g} {}^{\bar{r}}\tilde{w}_{\bar{g},ij}\bar{g}_i$ で計算される．ここで ${}^{\bar{r}}\tilde{w}_{\bar{g},ij}$ は ${}^{\bar{r}}w_{\bar{g},ij}$ から μX 原理を適用した次式によって計算される．

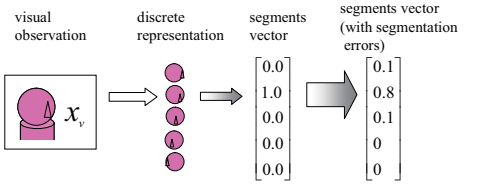


Fig.2 Example segment vectors of the discrete representation for gaze in the cases without and with errors in segmentation

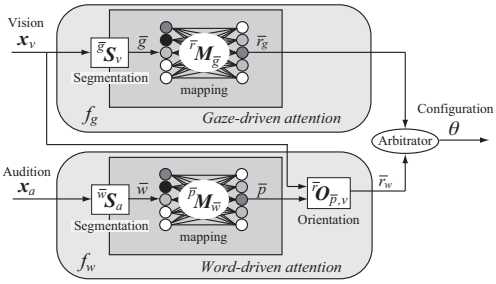


Fig.3 System of multimodal joint attention

れる．

$$ {}^{\bar{r}}\tilde{w}_{\bar{g},ij} = {}^{\bar{r}}w_{\bar{g},ij} \exp\left(-\frac{\sum_{k,k\neq i} {}^{\bar{r}}w_{\bar{g},kj}}{\alpha^2} \right) \qquad (1) $$

α はモジュール内の相互排他度の逆感度を表すパラメータである．この結果，j 番目の出力ノードと i 番目の入力ノードの相関が高く，かつ j 番目の出力ノードと他の入力ノードの相関が低いほど j 番目の出力ノードは活性化される．最後に出力 ${}^{\bar{r}}\bar{\boldsymbol{r}}_g$ はこの活性値を正規化した値 ${}^{\bar{r}}\bar{r}_{g,j} = {}^{\bar{r}}a_{\bar{g},j} / \sum_k {}^{\bar{r}}a_{\bar{g},k}$ として得られる．${}^{\bar{r}}\bar{r}_{g,j}$ は ${}^{\bar{r}}\bar{\boldsymbol{r}}_g$ の j 番目の要素であり，視線情報のみを考慮した場合の JA 実行に選ばれる尤度を表す．

2·1.2 言葉による注意

言葉による注意モジュールはセグメンテーション関数 ${}^{\bar{w}}\boldsymbol{S}_a$，マッピング関数 ${}^{\bar{p}}\boldsymbol{M}_{\bar{w}} : \Re^{M_w} \rightarrow \Re^{M_p}$ および配置関数 ${}^{\bar{r}}\boldsymbol{O}_{\bar{p},v}$ によって構成される（Fig.3 下部）．${}^{\bar{p}}\boldsymbol{M}_{\bar{w}}$ は観測された発話 $\bar{\boldsymbol{w}}$ から想起される視覚パターン $\bar{\boldsymbol{p}}$ を出力し，経験を通して重みが更新される．${}^{\bar{p}}\boldsymbol{M}_{\bar{w}}$ は ${}^{\bar{r}}\boldsymbol{M}_{\bar{g}}$ と同様のネットワークで構築される．$\bar{\boldsymbol{p}}$ は相互排他性に基づく式 (1) と同様のプロセスで $\bar{\boldsymbol{w}}$ から計算される．

最後に対象物の配置は ${}^{\bar{r}}\boldsymbol{O}_{\bar{p},v}$ によって特定され，視覚パターンに応じた注視位置 ${}^{\bar{r}}\bar{\boldsymbol{r}}_w$ が出力される．実際には ${}^{\bar{r}}\bar{\boldsymbol{r}}_w$ は次式のようなベクトルの表現を用いて実行される．

$$ {}^{\bar{r}}\bar{r}_{w,j} = \begin{cases} s_j & \text{if } s_j > 0 \\ (1 - \sum_j^{M_r} s_j)/N_s & \text{otherwise} \end{cases} \qquad (2) $$

ここで $s_j = \sum_k^{M_r} \bar{p}_k o_{jk}$ は j 番目の位置での視覚情報 $\boldsymbol{x}_v$ における言葉がもたらす顕著性を表す．o_{jk} はオブジェクトの k 番目の視覚パターンが存在する場合 1，その他は 0 となる．また $N_s = M_r - \sum_k^{M_r} o_k$ である．${}^{\bar{r}}\bar{r}_{w,j}$ は ${}^{\bar{r}}\bar{\boldsymbol{r}}_w$ の j 番目の要素であり，音声情報のみを考慮した場合の JA 実行に選ばれる尤度を表す．

2·2 相互排他性に基づく統合と学習

2·2.1 統合

注視行動は各モジュールの μX 原理に基づいた出力 ${}^{\bar{r}}\bar{\boldsymbol{r}}_g$ と ${}^{\bar{r}}\bar{\boldsymbol{r}}_w$ を統合して行われる．統合されたベクトル $\bar{\boldsymbol{r}}$ は

$$ \bar{\boldsymbol{r}} = \frac{\sum_{k=g,w} \exp(\pi\mu_k)\bar{\boldsymbol{r}}_k}{\sum_{k=g,w} \exp(\pi\mu_k)} \qquad (3) $$

で計算される．ここで μ_g および μ_w はそれぞれ $\mu_g = \max_j\{\bar{r}_{g,j}\}$，$\mu_w = \max_j\{\bar{r}_{w,j}\}$ で計算される．π はモ

ジュール間での相互排他度の感度を表すパラメータである．ただし $\bar{\boldsymbol{r}}_g$ および $\bar{\boldsymbol{r}}_w$ は各要素の総和が 1 になるように正規化されている．よって相互排他的な出力ほど大きな値をもち，選択されるようになる．最後に $\bar{\boldsymbol{r}}$ は $\theta_j = \bar{r}_j / \sum_k \bar{r}_k$ で正規化され，$\boldsymbol{\theta}$ は注視確率として用いられる．

2·2.2 学習

学習者は対象物を見つけたとき，各モジュールのマッピングの入力要素の最大値と注視行動に選択された出力の要素の結合を強める．この学習では学習者は常に養育者と同じ対象物を見て重み更新を行うわけではない．先行研究 [3][4][5] では養育者の視線方向あるいは発話された対象と一致する位置に対象物が頻繁に見つけられるという統計的制約を利用することで明示的な教示なしに学習が可能であることを示した．

本システムでもこの統計的制約からマッピングの入出力の相関を累積的に学習するアプローチを採用する．注視行動を行い対象物が発見できたとき，学習者は各マッピングの入力要素の最大値と注視行動に選択された出力との結合の重みを Δ 増やすと同時にそれ以外の出力との結合の重みは側抑制として Δ_l 減らす．

2·3 セグメンテーションのエラーモデル

本システムでは離散表現のための入力のセグメンテーション能力を事前に与えている．しかし，実世界では認識誤差による入力のセグメンテーションエラーを避けることはできない．そのため，セグメンテーションエラーはシステムに組み込まれる必要性がある．

本システムではセグメンテーションに対するエラーモデルを次式によって定義する．

$$e_{g,ij} = \frac{1}{\sqrt{2\pi}\sigma_g} \exp\left(-\frac{(i-j)^2}{2\sigma_g^2}\right) \qquad (4)$$

ここで σ_g はエラー率を調整するパラメータである．$e_{g,ij}$ は j 番目の要素が入力された場合の入力ベクトル $\bar{\boldsymbol{g}}$ の要素 i の尤度を表す．すなわち，近隣の入力ほど誤認識しやすいエラーモデルとなっている（Fig.2 参照）．$e_{w,ij}$ も同様の式で表される．本稿のシミュレーションでは各マッピングにおいて 5 割の入力をエラーのない識別容易な入力とし，残り 5 割の入力を 50% のエラーをもつ識別の難しい入力としている．

3. マルチモーダル発達シミュレーション

マルチモーダルな共同注意の発達過程における μX 原理の有効性を示すために計算機シミュレーションを行った．実験 1 では μX 原理による相互促進的学習の効果を検証した．実験 2 では養育者の行動を変えた場合におけるマルチモーダルな共同注意の発達過程の差異を検証した．これらのシミュレーションの基本設定を説明した後，結果を順次示していく．

3·1 基本設定

ラベルを教示するための養育者-幼児間インタラクションのシミュレーションは以下に示す設定で行われる．学習者と養育者は互いに向かい合っており，その間に 100 等分されたテーブルが存在する．対象物は 100 個準備され，そのうち 10 個の対象物がランダムに選択され，テーブル上のいずれかの位置に重複しないように配置される．テーブル上の対象物は 10 試行毎に入れ替えられる．1 試行毎に養育者は環境中から 1 つの対象物を見て，そのラベルを発話する．対象物は固有のラベルをもつ．学習者は注意モジュールを利用して養育者の意図している対象物の場所を見ようとする．最後に学習者は JA 成功の是非に関わらず，対象物を見た場合に各注意モジュールの重みを更新する．このインタラクションを繰り返す．

重みの各パラメータはそれぞれ $\Delta = 1.0, \Delta_l = 0.01$ としている．またモジュール内およびモジュール間の相互排他度を表す各パラメータはそれぞれ $\alpha = 1.0, \pi = 25.0$ である．これらのパラメータは経験的に決定した．実験 1 では養育者の対象物選択はランダムであるが，実験 2 では幼児の学習に対応して選択を変化させる．

3·2 実験 1：相互促進学習

各モジュールの学習性能における μX 原理の効果を評価するため，100,000 ステップのインタラクションを 20 回行った．Fig.4 は JA の成功率の軌跡を示し，各データポイントおよび分散は最後の 1,000 ステップを考慮して計算されている．

この結果から μX 原理を式 (1) により出力のみ適用した場合 (*intra-μX multimodal*：▲) が μX 原理の適用なし (*non-μX multimodal*：$*$) に比べ JA 成功率上昇が速いことから，モジュール内の μX 原理が学習速度に寄与しているといえる．また μX 原理を式 (3) により統合のみに適用した場合 (*inter-μX multimodal*：△) と適用なし (*non-μX multimodal*) を比べることでモジュール間の μX 原理が JA 成功率に寄与しているといえる．そして出力および統合の両方に適用した場合 (*double-μX multimodal*：○) のパフォーマンスが最も良くなってい

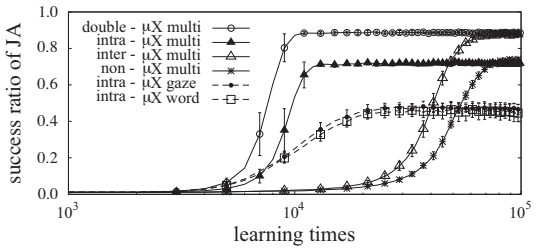
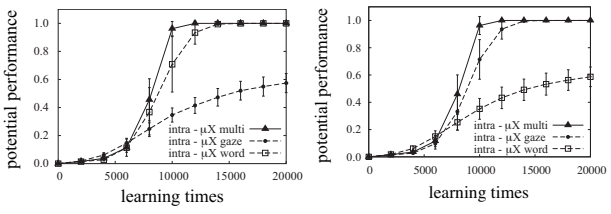


Fig.4 Success rate of joint attention



(a) Gaze-to-location mapping　(b) Word-to-object mapping

Fig.5 Potential performances of (a) the gaze-to-location mapping and (b) the word-to-object mapping

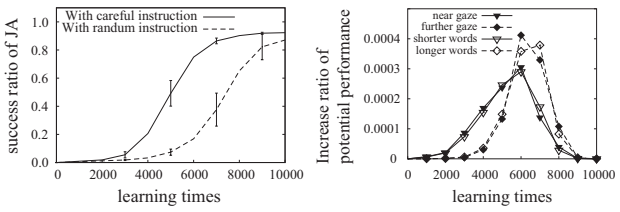
ることがわかる．

視線による注意 (*intra-μX gaze*:●)，言葉による注意 (*intra-μX word*:□) モジュールのみを利用して JA を学習する場合，JA の成功率はマルチモーダルシステムに比べ低い．Fig.5(b) および Fig.5(c) は 1,000 ステップ毎に各モジュールのマッピングの学習率を評価したグラフである．一方のモダリティのみを利用した学習では利用したモダリティの学習性能 (Fig.5(b) の●および Fig.5(c) の□) の学習は不十分であるが，もう片方のモダリティの学習性能 (Fig.5(b)) およびマルチモーダルの学習性能 (Fig.5(c) の○) は十分な学習ができている．これは一方のマッピングがもう一方のマッピングを修正するという結果であり，マルチモーダルシステムによってマッピング間で相互促進効果がもたらされることを示した．

3·3 実験 2：養育者を考慮した発達

次に “Learning from Easy Mission” メカニズム [9] に基づく養育者の行動が学習者の学習にもたらす促進作用について検証する．初めにより近い場所にあり，より簡単なラベルをもつ対象物を教えようとする行動は人の養育者と共通し，より最もらしい発達過程をもたらすと考えられる．本稿のシミュレーションでは養育者は各視線および各発話による学習者との JA を評価し，学習者が JA 可能な対象に対して教えるというような学習者の視線追従および語彙理解の能力に配慮した振る舞いをとる．学習者は *double-μX multimodal* に固定されている．

Fig.6(a) では養育者の行動が異なるケースでの学習過程を比較した．学習者の発達を配慮した行動 (実線) が配慮していない行動 (破線) に比べ，JA の成功率上昇が速いことを示している．すなわち，養育者の幼児の発達を考慮した振る舞いが幼児の学習を加速させる要因と成り得ることを示した．また Fig.6(b) では養育者が配慮した行動を行う条件下での近くの視線 (▼の実線) と遠くの視線 (◆の破線) の視線による注意モジュール，短い言葉 (▽の実線) と長い言葉 (◇の破線) の言葉による注意モジュールの学習性能の上昇率を示す．実線はエラーなしの認識の簡単な入力群を表し，破線はエラー率 50% の認識の難しい入力群を表す．このグラフでは難しいグループ (Fig.6(b) の破線) の学習性能は簡単なグループ (Fig.6(b) の実線) の学習性能上昇後に起きており，難しいグループの学習は簡単なグループの学習により促進されるという促進作用が示されている．



(a) Success rate of JA　(b) Potential performance

Fig.6 Success rate under careful target choices: (a) comparison to that under random choice and (b) increase ratios of potential performance for the inputs from the easy and difficult groups

4. 結論

一連のシミュレーションによって，提案された μX 原理がマルチモーダル共同注意での同時学習において相互促進作用をもたらすことを実証した．そして μX 原理に基づくシステムにより，幼児の共同注意発達過程のある側面を再現できたと考えられる．視線による注意モジュールの曲線 (Fig.6(b)：▼の実線および◆の破線) での視線追従可能な領域の拡大は 12ヶ月から 18ヶ月までの幼児の発達でも観測されている [1]．一方，視線追従可能な領域の拡張に伴う視線および言葉による注意モジュール間の相互促進作用は 18ヶ月までに言葉の学習に親の視線情報を利用するようになるという知見と一致している [2]．従って，より近い領域から視線追従可能になることと一致し，親の配慮された教示行動によって促進されている可能性があることを示した．

本稿ではセグメンテーション能力は事前に与えられ，学習すべき情報が事前に決められていた．しかし，実環境ではマッピングの学習精度を上げるために，セグメンテーション能力のパラメータを変えることで学習対象の情報の次元および解像度を適宜調整する必要性があると考えられる．ゆえに我々は次にマッピングと同時にセグメンテーションの問題を解くことで，それらの同時学習がもたらす相互促進作用を明らかにしていきたいと考えている．

[1] Butterworth, G., Jarrett, N.: “What minds have in common is space: Spatial mechamisms serving joint visual attention in infancy”, British Journal of Developmental Psychology, 9, 1, pp.55-72, 1991.

[2] Baldwin, D. A.: “Infants’ contribution to the achievement of joint reference”, Child Development, 62, 5, pp.875-890, 1991.

[3] Nagai, Y., Hosoda, K., Morita, A., Asada, M.: “A constructive model for the development of joint attention”, Connection Science, 15, 4, pp.211-229, 2003.

[4] Teuscher, C., Triesch, J.: “To care or not to care: Analyzing the caregiver in a computational gaze following framework”, In Proc. of the Third Intl. Conf. on Development and Learning (ICDL’04), California, USA, October, 2004.

[5] Roy, D., Pentland, A.: “Learning words from sights and sounds: a computational model”, Cognitive Science, 26, 1, pp.113-146, 2002.

[6] Yu, C., Ballard, H. D.: “Exploring the Role of Attention in Modeling Embodied Language Acquisition”, Fifth International Conference on Cognitive Modeling (ICCM’03), Bamberg, Germany, April, 2003.

[7] 菊池匡晃，荻野正樹，浅田稔：“顕著性に基づくロボットの能動的語彙獲得”, 日本ロボット学会誌, 26, 3, pp.226-270, 2008.

[8] Markman. E., Wachtel, G.: “Children’s use of mutual exclusivity to constrain the meanings of words”, Cognitive Psychology, 20, pp.121-157, 1988.

[9] Asada, M., Noda, S., Tawaratsumida, S., Hosoda, K.: “Vision-based reinforcement liearning for purposive behavior acquisition”, In Proc. of IEEE International Conference on Robotics and Automation, Nagoya, Japan, May, 1995.