マルチモジュールの相互促進的学習による共同注意発達過程の構成

Synthesizing a developmental process of joint attention through mutually facilitative learning of multi modules

○学　中野　吏（阪大院/JST ERATO）　　一般　吉川雄一郎（JST ERATO）

正　浅田　稔（阪大院/JST ERATO）　　一般　石黒浩（阪大院/JST ERATO）

Tsukasa NAKANO, Osaka Univ / JST ERATO, tsukasa.nakano@ams.eng.osaka-u.ac.jp

Yuichiro YOSHIKAWA, JST ERATO

Minoru ASADA, Osaka Univ / JST ERATO

Hiroshi ISHIGURO, Osaka Univ / JST ERATO

In this paper, we extend our previous model[7] for simultaneous learning of multi-modules for joint attention: gaze- driven attention and word-driven attention. Inspired from child language acquisition, mutually exclusivity bias is utilized for mutual facilitative learning in an inter-module manner by extending a modified Hebbian learning rule. We applied the proposed system to the computer simulation in more plausible setting as infant model and argued possible correspondence to some knowledge in developmental psychology.

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

1. はじめに

他者とのコミュニケーションのためには，他者と共同注意をすることが重要である．共同注意とは他者と同一の対象物に注意を向ける行動であり，複数の情報（相手の視線，言葉，指差し等）を手がかりに達成することができる．人の幼児は，これらのうち視線追従能力 [1] や語彙 [2] などを人の幼児は 2 歳頃までに段階的に獲得することが知られている．発達心理学の従来研究から，これらの能力の獲得過程には相互作用があることが伺える [3, 4] が，どのような学習メカニズムによりこれらの相互促進的な発達が可能になるのかについては明らかでない．

近年，人の発達過程を再現するロボットを構築することを通じて，このメカニズムを知ることを目指す構成論的アプローチが注目されており，視線追従能力 [5] や語彙 [6] をロボットに獲得させる研究が行われている．しかし，単一機能の学習に焦点を当てたものが多く，複数の機能の獲得過程における相互作用は考慮されていない．

そこで本研究では，前研究 [7] で提案した視線および言葉を手がかりとするマルチもジュールによる共同注意システム，すなわち視線追従能力および語彙のマッピングの同時学習メカニズムを拡張し，実際の親子のインタラクション場面により忠実な設定の計算機シミュレーションを実施し，幼児の認知発達過程における相互作用を構成論的に理解することを図る．本研究で提案する共同注意学習モデルは，前研究 [7] と同様に，この相互排他性バイアスに基づく学習メカニズムを語彙と視線追従の両方のマッピング学習に適用するものであり，各モダリティでの学習の効率化と同時に，モダリティを跨いだ相互促進の効果 [7] が期待される．

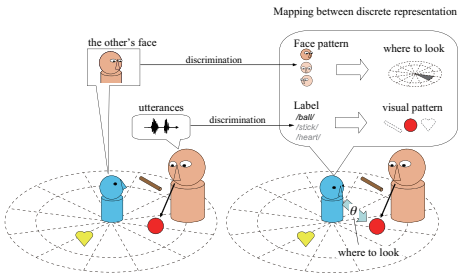


Fig. 1: Setting for learning multimodal joint attention

本稿では，まず次節で問題設定について述べ，前研究 [7] で提案したマルチモーダル共同注意システムの拡張したシステムについて述べる．3 節で，人の発達により忠実な設定の計算機シミュレーションについて述べ，提案手法の発達モデルとしての可能性を議論する．

2. 相互排他性原理に基づく共同注意

本研究で想定する学習者と養育者のインタラクションの状況を Fig.1 に示す．養育者は学習者の周囲に存在する複数の物体の中から一つを選択し，これを見ながら，その名称（以下ラベル）を発話する．学習者は Fig.2 に示す共同注意学習システムにより，観測した情報をもとに学習中の視線および言葉による注意モジュールを利用して物体を注視し，その経験から各注意モジュールのパラメータを更新する．このインタラクションを通じて，学習者は視線追従に

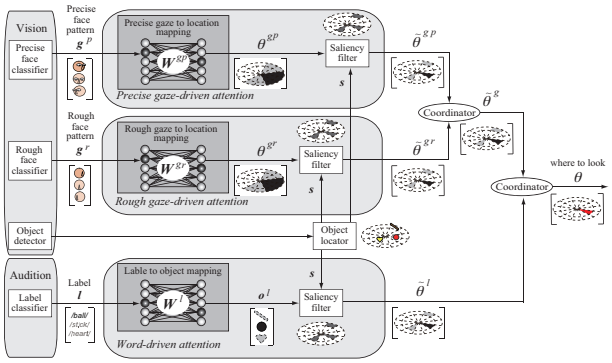


Fig. 2: Learning mechanism of joint attention with two modules: gaze-driven attentional one (top) and word-driven one (bottom)

必要な養育者の顔についての観測 $\boldsymbol{g}^r$ と $\boldsymbol{g}^p$ とその方向にあたるそれぞれの位置 $\boldsymbol{\theta}^{g^r}, \boldsymbol{\theta}^{g^p}$ へのマッピングをそれぞれ学習する．また語彙に必要なラベル $\boldsymbol{l}$ とラベルが指す物体 $\boldsymbol{o}$ のマッピングを学習する．ここで，$\boldsymbol{g}^r$ は顔の向きの違いが識別できる程度の粗い解像度での顔の観測情報であり，$\boldsymbol{g}^p$ は目の向きの違いまでも識別できる程度の高い解像度での顔の観測情報である．ここで $\boldsymbol{g}^r, \boldsymbol{g}^p, \boldsymbol{\theta}^{g^r}, \boldsymbol{\theta}^{g^p}, \boldsymbol{l}$ および $\boldsymbol{o}$ は予め登録された情報と観測情報との一致度を要素とし，最も一致した要素に 1，残りの要素に 0 が割り当てられるとし，それぞれ G^r, G^p, N, N, L および M 次元ベクトルであるとする．ただし，共同注意が成功したか否かの直接的な情報（教示や強化信号）は学習者に与えられないとする．その代わり学習者は様々な状況での経験を通して，共同注意が成功した場合の対応関係の学習頻度が他と比べ多いという統計的偏りを利用することで次第に正しいマッピングを学習する．

以下では，環境は N 個のスポットに分割され，教示される物体の候補は M 個とする．T ステップごとに M_o 個の物体がランダムに選択され，一つのスポットに重複しないように配置される．ここで 1 ステップとは，教示者が一つの物体を注視し，そのラベルを発話してから，ロボットがこれらの情報をもとに物体を注視し，各マッピングを更新するまでの間を指す．

2.1 視線による注意モジュール

視線による注意モジュールは，養育者の顔についての観測情報 $\boldsymbol{g}^r$ と $\boldsymbol{g}^p$ がそれぞれ入力されると，共同注意のために注視されるべき位置を

$$\boldsymbol{\theta}^{g^r} = \boldsymbol{W}^{g^r} \boldsymbol{g}^r \tag{1}$$

$$\boldsymbol{\theta}^{g^p} = \boldsymbol{W}^{g^p} \boldsymbol{g}^p \tag{2}$$

のように出力する．ここで，$\boldsymbol{\theta}^{g^r}$ および $\boldsymbol{\theta}^{g^p}$ は，養育者の視線を手がかりとしたときに環境中の各スポットが注視されるべき程度を表す．$\boldsymbol{W}^{g^r}, \boldsymbol{W}^{g^p}$ はそれぞれ $N \times G^r, N \times G^p$ 行列で表現されるマッピングの結合荷重である．ここで学習者は環境中のいずれかの物体を必ず見るとする．そのため，$\boldsymbol{\theta}^{g^r}$ の i 番目の要素が

$$\tilde{\theta}_i^{g^r} = \begin{cases} \theta_i^{g^r} & \text{if } s_i > 0 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

であるような N 次元ベクトル $\tilde{\theta}^{g^r}$ に修正される．同様に θ^{g^p} も $\tilde{\theta}^{g^p}$ に修正される．ここで $\boldsymbol{s}$ は環境中の物体配置を表現し，i 番目のスポットに物体が配置されていればその物体の ID である $id(> 0)$ が，そうでなければ 0 が i 番目の要素である s_i に設定される N 次元ベクトルである．

2.2 言葉による注意モジュール

言葉による注意モジュールは，養育者が発した物体のラベル $\boldsymbol{l}$ が観測されると，共同注意のために注視されるべき物体を

$$\boldsymbol{o}^l = \boldsymbol{W}^l \boldsymbol{l} \tag{4}$$

のように出力する．ここで $\boldsymbol{W}^l$ はマッピングのパラメータである $M \times L$ 行列である．

養育者の発話を手がかりとしたときの環境中の各スポットが注視されるべき程度を要素とする N 次元ベクトル $\tilde{\theta}^l$ は，その i 番目の要素を

$$\tilde{\theta}_i^l = \begin{cases} o_i^l & \text{if } s_i > 0 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

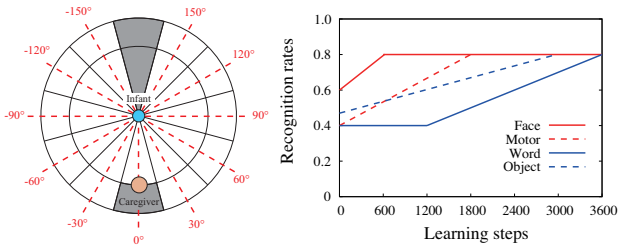
のように決定することで求められる．

2.3 統合

各注意モジュールの出力 $\tilde{\boldsymbol{\theta}}^{g^r}, \tilde{\boldsymbol{\theta}}^{g^p}$ と $\tilde{\boldsymbol{\theta}}^l$ は $\tilde{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}^{g^r} + \tilde{\boldsymbol{\theta}}^{g^p} + \tilde{\boldsymbol{\theta}}^l$ となるように統合され，$\tilde{\boldsymbol{\theta}}$ は正規化される．これをもとに注視するスポットが確率的に選択される．i 番目のスポットが選択されたとき，視線の先を表す $\boldsymbol{\theta}$ は i 番目の要素が 1，その他の要素が 0 となり，物体を表す $\boldsymbol{o}$ は i 番目のスポットに存在する物体の要素が 1，その他の要素が 0 となる．$\boldsymbol{\theta}$ と $\boldsymbol{o}$ は結合荷重の更新に用いられる．

2.4 相互排他性原理に基づく学習則

各試行において，学習者は観測情報 $\boldsymbol{g}^r, \boldsymbol{g}^p, \boldsymbol{l}$，と注視により得られた情報 $\boldsymbol{\theta}, \boldsymbol{o}$ をもとに各モジュールの学習を行う．学習者が注視経験により得た $\boldsymbol{\theta}$ と $\boldsymbol{o}$ は，養育者のものと必ずしも一致しない．しかし，対応が正しいときの経験が他と比べ頻度が高いことを利用した統計学習が可能であるこ



(a) Spots for joint attention

(b) Recognition rates

Fig. 3: (a)Environment of infant-caregiver interaction and (b) Recognition rates of inputs and outputs

とが示されている [5, 6]．また対応関係がどれだけ相互排他的に見えるかに応じて，観測された対応関係の学習率を調整する交差投錨型ヘッブ学習が提案されている [8]．本研究ではこれを修正した学習則をマルチモーダルな共同注意学習の問題に適用する．

学習者は自身の注視経験に基づいて各マッピングの結合荷重を更新する．ここでは $\boldsymbol{g}^r$ を入力とするマッピングの更新則についてのみ説明するが，他も同様である．$\boldsymbol{g}^r$ と出力 $\boldsymbol{\theta}$ の 1 である要素を，それぞれ勝者 i^* と j^* と呼ぶ．ここで t ステップ目にマッピング $\boldsymbol{W}^{g^r}$ における i^* 番目の入力要素と j^* 番目の出力要素の間の結合荷重 $w^{g^r}_{i^*j^*}(t)$ は

$$w^{g^r}_{i^*j^*}(t+1) = w^{g^r}_{i^*j^*}(t) + \eta^{g^r}_{i^*j^*}\eta^{\theta}_{i^*j^*}(g^r_{i^*}\theta_{j^*} - w^{g^r}_{i^*j^*}(t)) \quad (6)$$

のように更新される．ここで $\eta^{g^r}_{ij}$ と η^{θ}_{ij} はそれぞれマッピングの入力要素の排他度と出力要素の排他度を表す．入力要素の排他度は

$$\eta^{g^r}_{ij} = \exp\left(-\frac{\sum_{k,k\neq j} w^{g^r}_{ik}(t)}{\alpha^2}\right) \quad (7)$$

と計算される．η^{θ}_{ij} も同様にして計算される．ここで，α は相互排他性の逆感度を表すパラメータである．

同時に，入力の勝者要素と出力の勝者以外の要素との結合荷重は側抑制によって

$$w^{g^r}_{ij^*}(t+1) = w^{g^r}_{ij^*}(t) - \beta\eta^{g^r}_{ij^*}(g^r_i\theta_{j^*} - w^{g^r}_{ij^*}(t)) \quad (8)$$

のように減らされる．また出力の勝者要素と入力の勝者以外の要素との結合荷重も同様にして側抑制がかけられる．ここで β は側抑制の強さを表すパラメータである．以上の学習則によって学習者は養育者のもつ相互排他的なマッピングを学習する．

3. 幼児の発達過程再現シミュレーション

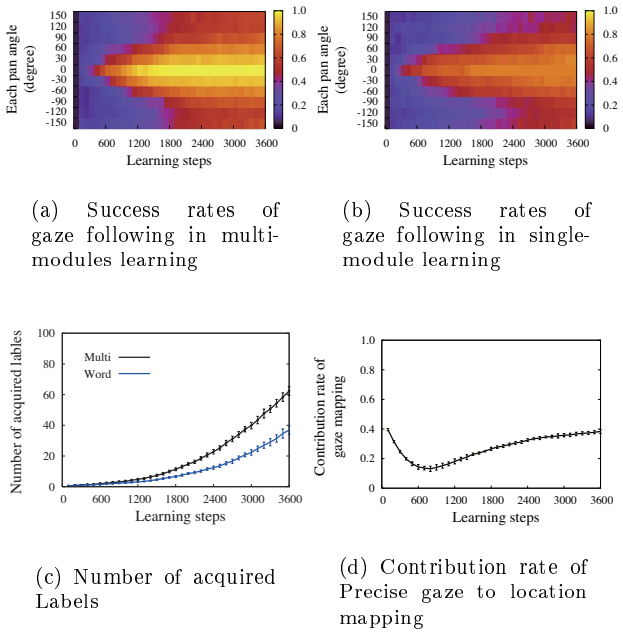
本稿では自然な親の行動と環境，および共同注意発達時期の幼児の識別能力を考慮したインタラクションを想定し，計算機シミュレーションを行った．実験設定の詳細について述べた後，計算機シミュレーションにおける提案システムの学習過程と幼児の発達過程との相同性について議論する．

学習者は Fig.3(a) に示す空間において，養育者との対面インタラクションを通じて共同注意を学習する．養育者は物体が配置されたスポットを注視し，その物体のラベルを発話する．学習者は二通りの解像度で観測された養育者の顔についての情報，および発話ラベルをもとに注視位置を決定する．ただし，養育者が常に注視スポットの方向をまっすぐ向いているとは限らない．そこで養育者の頭および目の方向は，養育者が注視すべきスポットへの方向を中心とするガウス分布にそれぞれ従うとし，0.8 の確率で注視すべきスポットに向かうものとなるようなパラメータを採用した．

また発達心理学において幼児は共同注意の発達と同時期の 6〜18ヶ月の間に共同注意学習に必要な入・出力情報の識別能力を発達させることが知られている [9, 10, 11, 12]．顔の識別能力には視力の発達 [9]，首振りの運動感覚の識別能力には姿勢の発達 [10]，言葉の識別能力には単語分節能力の発達 [11]，物体の識別能力には物体のカテゴリ化能力の発達 [12] をそれぞれ参考にし，Fig.3(b) に示す，ステップ数に応じて識別率を変更させることとした．

上記の実験設定で 3600 ステップの共同注意学習シミュレーションを 20 回実施した．また，Fig.3(a) に示すように，$N = 21$ とする．$\boldsymbol{g}^r$ の解像度は養育者の顔がどのスポットを向いているかを識別できる程度とし，$G^r = N = 21$ した．$\boldsymbol{g}^p$ の解像度はさらに養育者の目がどのスポットに向いているかまで識別できる程度とし，$G^p = 21 \times 21 = 441$ とした．また $\boldsymbol{l}$ は養育者が発しうる全てのラベルにそれぞれ対応する要素を持つとし，$L = M = 200$ とした．また提案手法の学習の各パラメータは $\alpha = 1.0, \beta = 0.1$ とし，経験的に決定した．

Fig.4(a) と (b) に 100 ステップ毎に 1000 回の視線追従テストを行ったときの学習者の視線追従成功率を色の濃さで示す．横軸はステップ数，縦軸はテストされた視線方向のパン角を表している．このテストの際には養育者の顔と目は必ず正しい方向を見ているものとした．これらの結果から複数機能を同時学習した場合（Fig.4(a)）に，視線追従のみ学習した場合（Fig.4(b)）と比べ，視線追従の学習が加速していることがわかる．また提案システムの（Fig.4(a)）では幼児の前方である 0 °から幼児の後方である ±150 °に向け，徐々に成功率が上昇していることが伺える．一方で人の幼児は，12ヶ月頃には，幼児自身の視野内にある物体に対するものに限られるのに対し，18ヶ月頃では目の前に物体がない場合でも視野外の物体を探しあて，視線追従ができるようになること [1] が知られており，本節のシミュレーションの結果はこれに対応する可能性がある．また Fig.4(c)



(a) Success rates of gaze following in multi-modules learning

(b) Success rates of gaze following in single-module learning

(c) Number of acquired Labels

(d) Contribution rate of Precise gaze to location mapping

Fig. 4: Developmental process of (a),(b) gaze following, (c) lexical acquisition, and (d) gaze following in more detail.

に 100 ステップ毎にそれぞれラベルを聞いてすべての対象物の中から対応の正しい物体を選べるかをテストし，正しく選択できた語彙の数の遷移を示す．Fig.4(c) から提案システムの結果（黒）が視線追従を利用しない語彙学習の結果（青）と比べ，獲得語彙数が多いことがわかる．幼児は 18ヶ月頃までに語彙学習に親の視線情報を利用していること [3] が知られており，この結果は語彙学習に視線情報を利用することで効率的な学習ができることに気づく可能性があることを示している．

最後に Fig.4(d) に 100 ステップ毎に 1000 回の視線追従テストを行ったときの注視したスポットの選択での解像度の高い注意モジュールの寄与率を示す．寄与率は $\tilde{\theta}^{g^p}_{i^*}/(\tilde{\theta}^{g^p}_{i^*}+\tilde{\theta}^{g^r}_{i^*})$ で求められる．この結果から 600 ステップ頃では解像度の粗い注意モジュールをよく利用しているのに対し，その後徐々に解像度の高い注意モジュールも利用するようになっていることがわかる．この結果は幼児が 15ヶ月頃から徐々に養育者の頭の方向だけでなく，目にも注目して視線追従を行うようになるという知見 [13] に対応する可能性がある．

4. 結言

そこで本研究では視線および言葉を手がかりとするマルチモジュールによる共同注意システムを，人の親子間インタラクションをより忠実な設定の計算機シミュレーションに適用し，人の幼児の発達過程との相同性を議論した．

幼児の発達過程再現シミュレーションにおいて提案手法が，Butterworth et al.[1] の示す視線追従の段階的発達過程を再現し，Baldwin[3] の示す 18ヶ月児が視線情報を利用した方が効率的に語彙学習できることに気づく可能性があることを示した．さらに視線追従に頭の方向だけでなく，目の方向も徐々に利用するようになるという発達過程 [13] も再現することができた．本研究の設定では入力および出力の情報を識別するための離散化能力は予め与えられていた．幼児の発達過程においては，視線追従および語彙のマッピングに必要な認識能力，すなわち離散化能力の発達がマッピング学習と同時進行していることが発達心理学の知見から伺える．従って，マッピングと入力情報の離散化の同時学習問題に取り組むことで，幼児の発達過程を支える学習メカニズムの構成論的理解を深化させることが重要な今後の課題であるといえる．

参考文献

[1] G. Butterworth and N. Jarrett: “What minds have in common is space: Spatial mechamisms serving joint visual attention in infancy”, British J. of Dev Psycho, 9(1), 55–72(1991).

[2] E. Bates et al.: “Individual differences and their implications for theories of language development”, The handbook of child lang, 96–151(1995).

[3] D.A. Baldwin: “Infants’ contribution to the achievement of joint reference”, Child Dev, 62(5), 875–890(1991).

[4] P.E. Spencer: “Looking Without Listening: Is Audition a Prerequisite for Normal Development of Visual Attention During Infancy?”, J. of Deaf Studies and Deaf Edu, 5(4), 291–302(2000).

[5] Nagai et al.: “A constructive model for the development of joint attention”, Connection Sci, 15(4), 211–229(2003).

[6] D.K. Roy and A.P. Pentland: “Learning words from sights and sounds: a computational model”, Cognitive Sci, 26(1), 113–146(2002).

[7] 中野ら：“相互排他性原理に基づくマルチモーダル共同注意”, 第 26 回日本ロボット学会学術講演会,(2009).

[8] Y. Yoshikawa et al.: “Unique association between self-occlusion and double-touching towards binding vision and touch”, Neurocomp, 70(13-15), 2234–2244(2007).

[9] AM. Norcia and CW. Tyler: “Spatial frequency sweep VEP: visual acuity during the first year of life. Vision research, 25(10), 1399–1408(1985).

[10] R. Illingworth: “Basic Developmental Screening: 0-2 years”, Oxford: Blackwell Scientific,1973.

[11] C.L. Steger and J.F. Werker: “Infants listen for more phonetic detail in speech perception than in word-learning tasks”, Nature, 388(6640),381–382(1997).

[12] B. Youger and L. Cohen: “Infant Perception of Correlations among Attributes”, Child Dev,54,858–867(1983).

[13] V. Corkum and C. Moore: “Development of joint visual attention in infants”, Joint Visual Attention. Erlbaum. Hillsdalle.NJ.61–83.