Proceedings of the 2005 JSME
Conference on Robotics and Mechatronics
2A1-N-043, June 2005

動き情報の利用による共同注意学習の高速化

Motion Information Accelerates Learning of Joint Attention

正　長井 志江（情報通信研究機構）

Yukie NAGAI, National Institute of Information and Communications Technology, yukie@nict.go.jp

Focusing on the evidence that humans utilize motion information of others' actions, I propose a model by which a robot learns to establish joint attention with a human by using both static and motion information of human actions. The static information, which is an edge image of a human face, provides the exact direction of the human gaze even though it is difficult to interpret. On the other hand, the motion information, which is detected as an optical flow when the human shifts his/her gaze, provides a rough but easily understandable relationship between the direction of the human gaze and the robot's motor command to follow it. These complementary characteristics enable a robot to efficiently acquire high performance of joint attention through natural interactions with a human. Experimental results show that the motion information accelerates the learning of joint attention while the static information improves the task performance.

Key Words: joint attention, learning, motion information, optical flow

1. 緒　言

本論文では，視覚上に検出される他者の動き情報を利用した共同注意の学習モデルを提案する．共同注意 [7] は他者の視線方向を追従して，他者と同一の対象物に注意を向ける行為であり，人間の認知発達の基礎として，また人間－ロボット間の社会的コミュニケーションの第一歩として注目を集めている（図 1 参照）．これまで，構成論的アプローチから，人間の認知発達メカニズムの理解と，その知見を応用した自律的学習ロボットの実現を目的として，いくつかの共同注意学習モデルが提案されてきた．著者 [9, 10] は，共同注意学習において学習者がもつべき埋め込み機能と，養育者の教示，そしてそれらの発達的要素がもたらす学習への効果について，さまざまな側面から議論してきた．その結果，学習者にとって，環境内の特徴的な対象（顔パターンや，動きや明るい色などの特徴をもつ対象物）へ注意を向けるという選好注視機能が，共同注意の学習に必要な養育者との対象物への同時注視という行動経験を産み，この経験から知覚－運動－環境変化間の随伴性を学習することで，共同注意能力が獲得されることを示した．また，この学習と並行して，学習者自身の知覚能力の発達や，学習者の能力に合わせた養育者の教示の適応的変化が起きることで，共同注意の学習性能が改善されることも示した．この結果をもとに，学習モデルを実ロボットに実装して，人間との自然なインタラクションを通して共同注意能力を獲得させる実験も試みられている [2, 8]．Triesch et al. [1, 4] は幼児発達研究者と協力し，幼児の様々な観察データに基づいて共同注意の計算論的学習モデルを提案した．彼らは，学習者がもつ特徴的な対象物に対する選好注視機能と，それを達成したときの報酬ベースの学習能力，そして，学習者と養育者との間で十分に同時注視を成立させ得る環境設定が，共同注意能力を獲得させるのに必要十分な要素であると提言した．しかし，これまでの研究は，いずれも学習者の知覚情報として，共同注意を行う相手の静的な姿勢情報や静止画像としての顔パターンを用いているのみであり，相手の視線変化時に検出される動き情報の役割については議論してこなかった．

これに対して，人間は視覚上に知覚される動きに非常に敏感であることが知られている．特に，幼児においてはそれが顕著であり，共同注意能力の発達においても，他者の視線変化時の動き情報が幼児の視線方向の認識を援助し，視線追従行動を促進していることが指摘されている [5, 6]．

本論文ではこれらの認知発達学的知見に基づき，他者の注視動作の静的画像情報に加え，視線変化時の動き情報を利用した共同注意の学習モデルを提案する．静的情報としては，他者が対象物を注視しているときの顔のエッジ画像を用いる．これは，他者の視線方向に関しての十分な情報を含んでおり，学習者が共同注意の知覚－運動パターンを学習するのに際し，正確な入出力関係を与えてくれると期待できる．これに対して，動き情報は，他者が対象物を注視するために視線を変化させたときの，顔画像領域のオプティカルフローとして検出される．オプティカルフローは，エッジ画像に比べて他者の視線方向を表す情報の精度は低いが，それ自体ですでに方向性を示しており，学習者の視線追従運動への変換が容易であると考えられる．またフローの示す方向が，多くの場合，学習者が追従すべき運動方向ときれいに対応していることから，知覚－運動パターンの学習を加速する効果をもつと期待できる．本研究ではこれら二種類の視覚情報を利用した学習モデルをロボットに実装し，検証実験を行うことで，共同注意発達におけるさまざまな知覚入力情報の役割と，人間の幼児におけるそれらの処理メカニズムの構成論的理解を試みる．

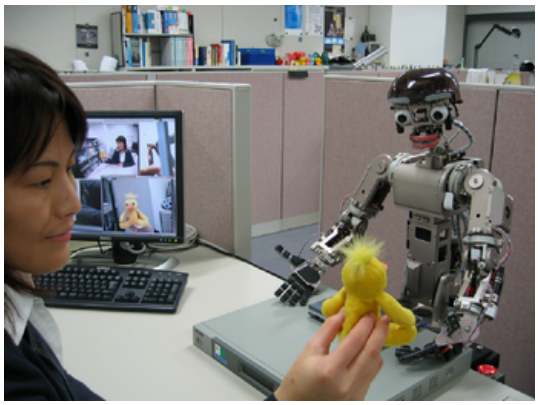


Fig. 1: Human-robot joint attention.

2. 動き情報を利用した共同注意の学習モデル

他者の視線変化時の動き情報と視線変化後の静的情報の両方を利用した，共同注意の学習モデルを図 2 に示す．本モデルは，画像特徴抽出器，視覚－運動学習器，運動出力調定器の 3 つのモジュールから構成される．ロボットは選好注視機能に基づく特徴的な対象物への同時注視経験を通して，本モデルにより人間との共同注意の能力を獲得する．

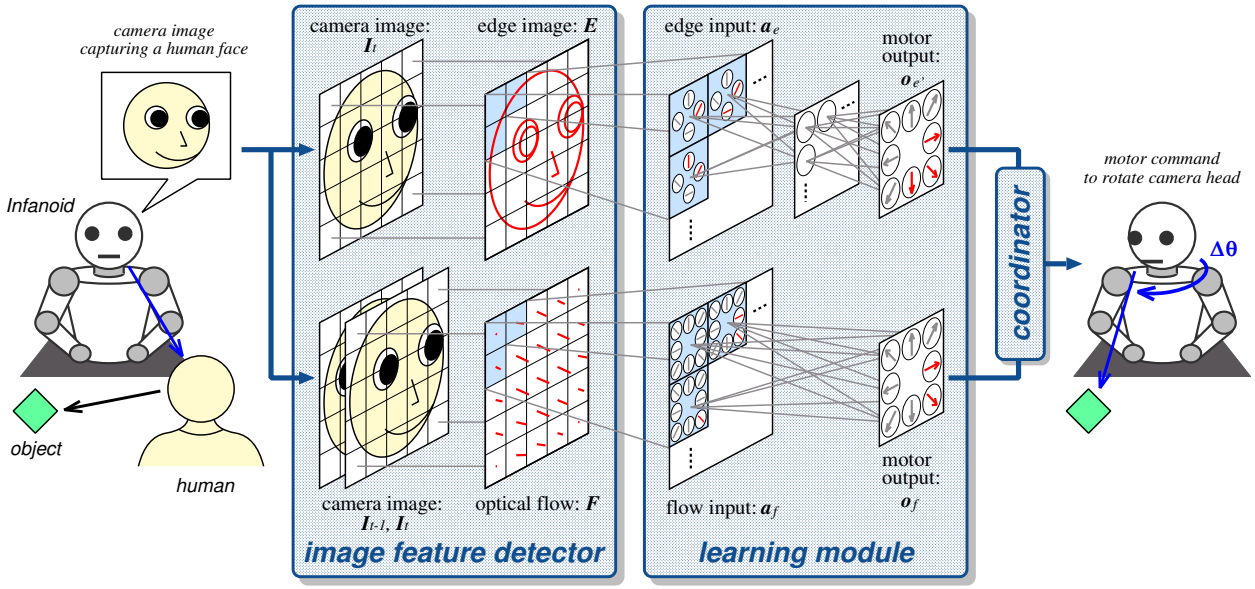


Fig. 2: A learning model for joint attention using the edge image of a human face as static information and the optical flow of the human's gaze shift as motion information.

2.1 画像特徴抽出器

画像特徴抽出器は，ロボットのカメラ画像 $\boldsymbol{I}_{t-1}$，$\boldsymbol{I}_t$ から，人間の注視動作時の顔のエッジ画像 $\boldsymbol{E}$ とオプティカルフロー $\boldsymbol{F}$ を抽出するモジュールである．図 3 (a) と (b) にロボットのカメラ画像の一例を示す．ロボットは (a) 周辺視カメラと (b) 中心視カメラをもち，(a) の画像で人間の顔をカメラ中心に捕らえた後，(b) の画像から本モジュールにより画像特徴 $\boldsymbol{E}$，$\boldsymbol{F}$ を抽出する．

まず，エッジ画像 $\boldsymbol{E}$ は方位選択性フィルタを用いて抽出する．4 方位 $(e_1,\ e_2,\ e_3,\ e_4) = (\text{—},\ \diagdown,\ |,\ \diagup)$ に反応選択性をもつ 4 つのフィルタで，カメラ画像 $\boldsymbol{I}_t$ から各方位のエッジ要素を集めた画像 $\boldsymbol{E}_n$ $(n = 1, \ldots,\ 4)$ を生成する．$\boldsymbol{E}_n$ の各画素 $(x,\ y)$ の値 $E_n(x,\ y)$ は，$\boldsymbol{I}_t$ の輝度値 $I(x,\ y)$ から微分フィルタを用いて以下のように計算される．

$$E_n(x,\ y) = \begin{cases} 1 & \text{if } \epsilon_n(x,\ y) > \epsilon_{\text{thr}} \\ 0 & \text{otherwise,} \end{cases}$$

$$\text{where } \epsilon_n(x,\ y) = \left| \sum_{i=-1}^{1} \sum_{j=-1}^{1} \alpha_n(i,\ j) I(x+i,\ y+j) \right| - \left| \sum_{i=-1}^{1} \sum_{j=-1}^{1} \beta_n(i,\ j) I(x+i,\ y+j) \right| \quad (1)$$

ここで，ϵ_{thr} は定数の閾値であり，係数 $\alpha_n(i,\ j)$，$\beta_n(i,\ j)$ は以下の値で与えられる．

$$\boldsymbol{\alpha}_1 = \boldsymbol{\beta}_3 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & -1 & -1 \end{bmatrix},\ \boldsymbol{\beta}_1 = \boldsymbol{\alpha}_3 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & 1 & -1 \end{bmatrix},$$

$$\boldsymbol{\alpha}_2 = \boldsymbol{\beta}_4 = \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 1 \\ 0 & -1 & 0 \end{bmatrix},\ \boldsymbol{\beta}_2 = \boldsymbol{\alpha}_4 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & -1 \\ 0 & -1 & 0 \end{bmatrix},$$

$$\text{where } \boldsymbol{\alpha}_n = \begin{bmatrix} \alpha_n(-1,-1) & \alpha_n(0,-1) & \alpha_n(1,-1) \\ \alpha_n(-1,\ 0) & \alpha_n(0,\ 0) & \alpha_n(1,\ 0) \\ \alpha_n(-1,\ 1) & \alpha_n(0,\ 1) & \alpha_n(1,\ 1) \end{bmatrix} \quad (2)$$

図 3 (c) に，(b) の矩形で囲まれた画像中心領域（168×168 画素）から抽出したエッジ画像 $\boldsymbol{E}$ を示す．これは，4 方位についてのエッジ画像 $\boldsymbol{E}_n$ を合成したもので，各方位性をもったエッジ (—, ＼, |, ／) を，それぞれ (赤, シアン, 青, 緑) で描いている．この静的入力情報としてのエッジ画像は，人間の視線方向を推定するのに十分な情報を含んでおり，ロボットはこれを用いることで，人間の視線を追従するための正確な運動出力を生成できると考えられる．

一方，オプティカルフロー $\boldsymbol{F}$ は，2 枚の連続するカメラ画像 $\boldsymbol{I}_{t-1}$，$\boldsymbol{I}_t$ 間の対応点検出によって求める．$\boldsymbol{I}_t$ の画像中心領域に対して，さらにそれを 24×24 画素の微小領域（受容野）に分割し，各受容野ごとに $\boldsymbol{I}_{t-1}$ との間でテンプレートマッチングによる対応点検出を行う．$\boldsymbol{I}_t$ における k 番目の受容野の中心位置を $(x_k,\ y_k)$，$\boldsymbol{I}_{t-1}$ における対応点を $(px_k,\ py_k)$ とすると，フロー $\boldsymbol{F}^k$ は 10 フレームの累積和として，

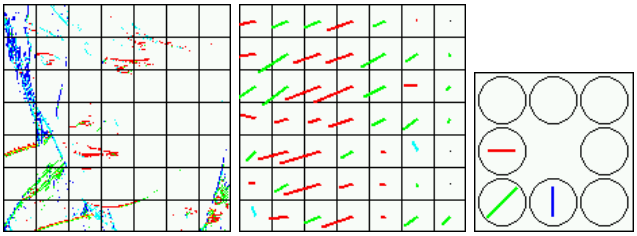
$$\boldsymbol{F}^k = \begin{bmatrix} \sum^{10\text{frames}} (x_k - px_k) \\ \sum^{10\text{frames}} (y_k - py_k) \end{bmatrix} \quad (3)$$

と求められる．図 3 (d) に，(b) の画像中心領域からのフロー検出結果を示す．これは，人間がロボットのカメラを注視している状態から (a) 内に観察される黄色い対象物の方向へ視線を移したときのオプティカルフローである．エッジの場合と同様に，方向によって 4 色に色分けされている．オプティカルフローはエッジ画像に比べて，人間の視線方向を推定するのにおおまかな情報しか含んでいないが，それ自体ですでに方向性を示しており，ロボットの視線追従運動への変換が容易であると考えられる．また (d) に見られるように，フローの示す方向は，多くの場合，ロボットが追従すべき運動方向ときれいに対応していることから，これを入力情報として利用することでロボットは共同注意の学習を加速させることが期待できる．

2.2 視覚－運動学習器

視覚－運動学習器は 2 つのニューラルネットワーク（NN）から構成され，エッジ入力と運動出力，そしてフロー入力と運動出力との関係を独立に学習するモジュールである（図 2 参照）．エッジ入力を処理する NN（エッジ NN）は，エッジ画像が人間の複雑な顔特徴をそのまま含んでおり，そこからロボットの視線追従運動を生成するまでの処理が困難であると予想されることから，

(a) peripheral camera image　(b) foveal camera image: $\boldsymbol{I}_t$



(c) edge image: $\boldsymbol{E}$　(d) optical flow: $\boldsymbol{F}$　(e) motor output: $\boldsymbol{o}_{e'f}$

Fig. 3: An example of input-output datasets.

3 層構造の NN を採用する．これに対して，オプティカルフロー入力を処理する NN（フロー NN）は，フローがすでに抽象化された情報であり，フローと人間の視線を追従するロボットの運動方向との対応関係が明瞭であることから，2 層の NN を用いて学習の高速化を試みる．

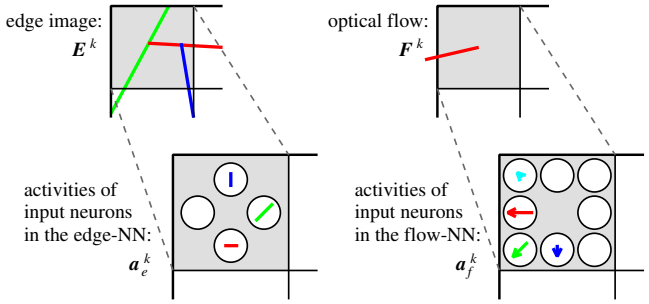
画像特徴を NN の入力にコーディングする方法として，まずエッジ入力は，4 方位 $(e_1, \ldots, e_4) = (—, \ldots, ／)$ に反応選択性をもつ 4 つのニューロンの活性度として表現する．図 4 (a) に，一受容野におけるエッジ画像から方位選択性ニューロンへのコーディングを示す．上図がエッジ画像，下図がそれをコーディングした結果である．コーディング後の図で，円が 1 つのニューロンに対応しており，円内の線分の向きがそのニューロンの選択的方位，長さが活性度を表している．k 番目の受容野における方位選択性ニューロンの活性度 $a^k_{e_n}$ $(n = 1, \ldots, 4)$ は，その受容野に含まれるエッジ量 E^k_n と全受容野での最大エッジ量との相対比から，

$$a^k_{e_n} = E^k_n / \max_k \sum_{m=1}^{4} E^k_m$$
$$\text{where} \quad E^k_n = \sum_{x_k} \sum_{y_k} E_n(x,\ y) \tag{4}$$

と計算される．また，オプティカルフロー入力は，8 方向 $(f_1,\ f_2, \ldots,\ f_8) = (←,\ ↖, \ldots,\ ↙)$ に反応選択性をもつ 8 つのニューロンの活性度として NN にコーディングされる．図 4 (b) にそのコーディング例を示す．k 番目の受容野における方向選択性ニューロンの活性度 $a^k_{f_n}$ $(n = 1, \ldots, 8)$ は，その受容野におけるフロー成分 $\boldsymbol{F}^k$ と 8 方向へのユニットベクトル $\boldsymbol{u}_n$ との内積を用いて，

$$a^k_{f_n} = \begin{cases} \boldsymbol{F}^k \cdot \boldsymbol{u}_n / \max_k \|\boldsymbol{F}^k\| & \text{if} \quad \boldsymbol{F}^k \cdot \boldsymbol{u}_n \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

と計算される．エッジの場合と同様に，図 4 (b) でも各ニューロンの活性度が 8 つの円内の矢印の長さとして表現されている．一



(a) the encoding of edge input　(b) the encoding of flow input

Fig. 4: The encoding of detected image features into the 4-orientations/8-directions selective neurons.

方，運動出力は，エッジ NN とフロー NN ともに 8 つの運動方向 $(e'_1, \ldots,\ e'_8) = (f_1, \ldots,\ f_8) = (←, \ldots,\ ↙)$ に反応選択性をもつ 8 つのニューロンの活性パターン $o_{e'_n}$，o_{f_n} として表現される．図 3 (e) に出力パターンの一例を示す．各ニューロンの表現はオプティカルフローのコーディング後の表現と同様である．これが次の運動出力調定器で，ロボットのモータコマンドに変換される．

以上の構造をもつ NN を，ロボットは選好注視機能に基づく人間との同時注視経験を通して学習する．選好注視を行った際の入出力関係を利用し，それを教師信号とした誤差逆伝搬法による学習を行うことで，共同注意のための知覚－運動パターンを獲得する．学習の仕組みについては文献 [10] を参照されたい．

2.3　運動出力調定器

運動出力調定器はエッジ NN とフロー NN からの出力を合成して，ロボットの視線追従運動のためのモータコマンド $\boldsymbol{\Delta\theta}$ を生成するモジュールである．エッジ NN からの出力 $o_{e'_n}$ とフロー NN からの出力 o_{f_n} の平均値を $o_{e'f_n}$ とすると，$\boldsymbol{\Delta\theta}$ は $o_{e'f_n}$ とユニットベクトル $\boldsymbol{u}_n$ の水平成分 u_{n_x}，垂直成分 u_{n_y} との積をとって，

$$\boldsymbol{\Delta\theta} = \begin{bmatrix} \Delta\theta_{pan} \\ \Delta\theta_{tilt} \end{bmatrix} = \begin{bmatrix} g_{pan} \sum_n u_{n_x} o_{e'f_n} \\ g_{tilt} \sum_n u_{n_y} o_{e'f_n} \end{bmatrix} \tag{6}$$

と求められる．ここで，g_{pan} と g_{tilt} はスカラーゲインである．ロボットの運動出力は，パン方向とチルト方向への視線変化の回転角度ベクトルとして表現される．

3.　実　験

3.1　実験設定

学習モデルを図 1 に示した幼児型ロボット Infanoid [3] に実装して，学習実験を行った．Infanoid は頭部にステレオ視をもち，目の 3 自由度（パン 2 自由度，チルト 1 自由度）と首の 3 自由度（パン 1 自由度，チルト 2 自由度）を用いて，視線方向を変化させることができる．また，左右の目にはそれぞれ周辺視カメラと中心視カメラの 2 種類の CCD カメラをもち，図 3 に示したような画像を獲得することができる．本実験では，視覚入力獲得のために左目の 2 つのカメラを用い，視線追従運動のために首の 3 自由度を用いた．チルト方向の 2 自由度については，式 (6) で求めた運動出力 $\Delta\theta_{tilt}$ を等分して 2 自由度に振り分けた．このようなロボットに対して，人間は正面に座り，試行ごとに注視対象物の位置を変化させて，ロボットの顔を注視した後に対象物へ視線を向けるという動作を繰り返し行った．

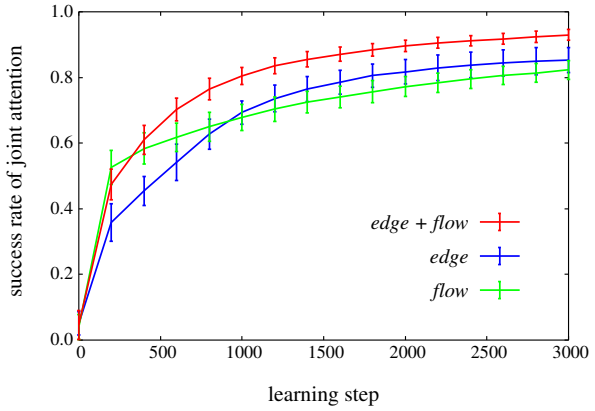


Fig. 5: The change in the task performance of joint attention over the learning period.

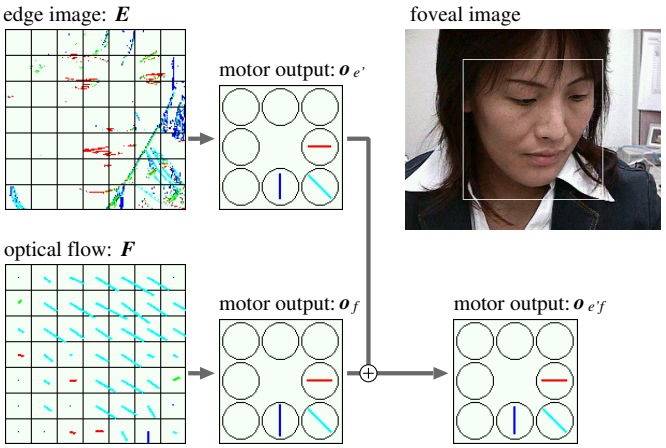


Fig. 6: The input-output dataset when the robot attempted to establish joint attention with the human who shifted her gaze from looking at the robot to looking at an object in the outer lower right of the foveal image.

3.2 学習における動き情報の役割についての評価実験

共同注意の学習における，他者の動き情報と静的情報の役割を評価する実験を行った．実験に際して，まず実環境で Infanoid に人間との同時注視を経験させ，学習に用いる 200 組の入出力データを獲得した．その後，獲得したデータを用いてオフラインでの学習実験を行った．

図 5 に学習回数に対する共同注意の成功率の変化を示す．これは，運動出力の角度誤差が ±8 度以下（中心視で対象物を捕らえられる範囲）であるときを共同注意の成功とみなした場合の，パフォーマンスの変化を示している．グラフ中，赤線がエッジ入力とフロー入力の両方を用いて学習を行った場合の実験結果，青と緑がエッジ入力もしくはフロー入力のどちらか一方のみを用いた場合の実験結果である．それぞれ，条件を変化させた 50 回の学習実験の平均と分散を示している．これらの比較より，(a) フロー入力は学習初期の立上り時間を加速させる効果があること，また，(b) エッジ入力は学習の立上りは遅いが，パフォーマンスの面で徐々にそれを改善していく効果があることが確認できる．そして，これらの 2 つの入力情報を同時に用いることで，学習の立上りが速く，かつ精度の高い共同注意能力を獲得できていることが分かる．

3.3 学習後の共同注意実験

エッジ入力とフロー入力の両方を用いた学習で獲得したモデルを Infanoid に実装して，人間との共同注意実験を行った．図 6 に，実験を行ったときの Infanoid が獲得した画像データとそこから抽出したエッジ，フロー情報を，そしてこれらの入力情報をもとに各 NN が出力したロボットの視線運動ベクトルを示す．ここでは，人間はロボットを注視している状態から画面右下にある対象物の方向へと視線を移動させている．この結果から，人間が注視している方向と，各 NN から出力されたロボットの運動ベクトル方向，そして合成された最終運動ベクトル方向とがきれいに対応しており，NN が適切な入出力関係を獲得できていることが確認できる．実際に，実環境で共同注意を行った結果は，学習時と同一人物，対象物の位置変化という条件で，成功率 90%（18/20 試行）であった．

4. 結 言

本論文では，視覚上に検出される他者の動き情報を利用した共同注意の学習モデルを提案し，実ロボットを用いた検証実験の結果を示した．認知発達学の分野では，他者の動き情報が幼児の視線認識や意図理解を助け，さまざまな認知機能の発達を促していることが示唆されている．本論文では，共同注意タスクにおいて，他者の視線変化時の動きをオプティカルフローとして検出し，それと学習者自身の視線追従運動ベクトルとの対応関係を獲得させることで，他者の注視動作の静的な情報のみを用いた場合よりも，学習を加速させる効果があることを示した．この結果は，幼児が外界の動きをどのように認識し，そして自身の認知機能の学習に役立てているのかを理解する上で，実証的なサポートとなる．今後は，学習実験そのものを動的な実環境の中で行うことで，環境全体からの入力としての動き情報の役割，そして他のさまざまな知覚情報の認知機能発達における役割について議論していく．

文 献

[1] E. Carlson and J. Triesch. A computational model of the emergence of gaze following. In *Proceedings of the 8th Neural Computation and Psychology Workshop*, 2003.

[2] K. Hosoda, H. Sumioka, A. Morita, and M. Asada. Acquisition of human-robot joint attention through real-time natural interaction. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2867–2872, 2004.

[3] H. Kozima. Infanoid: A babybot that explores the social environment. In K. Dautenhahn, A. H. Bond, L. Canamero, and B. Edmonds, editors, *Socially Intelligent Agents: Creating Relationships with Computers and Robots*, chapter 19, pages 157–164. Amsterdam: Kluwer Academic Publishers, 2002.

[4] B. Lau and J. Triesch. Learning gaze following in space: a computational model. In *Proceedings of the Third International Conference on Development and Learning*, 2004.

[5] J. D. Lempers. Young children's production and comprehension of nonverbal deictic behaviors. *The Journal of Genetic Psychology*, 135:93–102, 1979.

[6] C. Moore, M. Angelopoulos, and P. Bennett. The role of movement in the development of joint visual attention. *Infant Behavior and Development*, 20(1):83–92, 1997.

[7] C. Moore and P. J. Dunham, editors. *Joint Attention: Its Origins and Role in Development*. Lawrence Erlbaum Associates, 1995.

[8] A. Morita, Y. Yoshikawa, K. Hosoda, and M. Asada. Joint attention with strangers based on generalization through joint attention with caregivers. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3744–3749, 2004.

[9] Y. Nagai, M. Asada, and K. Hosoda. Developmental learning model for joint attention. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 932–937, 2002.

[10] Y. Nagai, K. Hosoda, A. Morita, and M. Asada. A constructive model for the development of joint attention. *Connection Science*, 15(4):211–229, 2003.